Stora Tentan in Bioinformatics

Reading guide for the oral PhD exam at DBB, Stockholm University

Arne Elofsson

2026-10-08

Table of contents

About this guide 8
What the exam is . . . 8
Programming and reproducible research . . . 9
Uncertainty: many numbers, different meanings . . . 11
Information theory in brief . . . 12
How each chapter is organised . . . 13
Topics . . . 14
Earlier versions of the instructions . . . 14
Additional material . . . 15
Companion textbook and contributors . . . 15
Licence . . . 15

I Topics 16

1 Sequence searching 17
1.1 Overview . . . 17
1.2 Homology and sequence similarity . . . 17
1.3 Dynamic programming and sequence alignment . . . 24
1.4 Heuristic search methods . . . 36
1.5 Alignment-free sequence comparison . . . 49
1.6 Profiles and multiple sequence alignments . . . 51
1.7 RNA homology search . . . 74
1.8 Whole-genome alignment and conservation . . . 78
1.9 Databases and domains . . . 81
1.10 Protein language models . . . 90
1.11 Benchmarking search methods . . . 98
1.12 Key concepts . . . 100
1.13 Recommended reading . . . 101
1.14 Background reading (open access) . . . 102
1.15 Student recommendations . . . 103
1.16 Review problems . . . 103

2 Databases 112
2.1 Overview . . . 112
2.2 How biological databases work . . . 112
2.3 Sequence databases . . . 119
2.4 Genome databases and browsers . . . 121
2.5 Structure databases . . . 123
2.6 Function and pathways . . . 124
2.7 Interaction databases . . . 126
2.8 Expression and proteomics repositories . . . 127
2.9 Medical and variant data . . . 128
2.10 Key concepts . . . 129

2.11 Recommended reading . . . 130
2.12 Background reading (open access) . . . 131
2.13 Student recommendations . . . 131
2.14 Review problems . . . 132

3 Evolution **139**
3.1 Overview . . . 139
3.2 The origin of life . . . 139
3.3 Basic evolutionary concepts . . . 142
3.4 Homologues, orthologues and gene duplication . . . 145
3.5 Protein evolution . . . 150
3.6 Detecting selection . . . 153
3.7 Genome evolution . . . 155
3.8 Key concepts . . . 157
3.9 Recommended reading . . . 158
3.10 Background reading (open access) . . . 158
3.11 Student recommendations . . . 160
3.12 Review problems . . . 160

4 Phylogeny **167**
4.1 Overview . . . 167
4.2 Basic definitions . . . 167
4.3 Phylogenetic methods . . . 170
4.4 Alignments, genomes and conflicting trees . . . 177
4.5 The tree of life . . . 182
4.6 Key concepts . . . 185
4.7 Recommended reading . . . 186
4.8 Background reading (open access) . . . 187
4.9 Student recommendations . . . 188
4.10 Review problems . . . 188

5 Genomics **196**
5.1 Overview . . . 196
5.2 Sequencing methods . . . 196
5.3 Genome assembly . . . 200
5.4 Mapping and variant calling . . . 204
5.5 Regulatory and epigenomic sequencing data . . . 211
5.6 Gene prediction . . . 213
5.7 Metagenomics . . . 217
5.8 Human genetics: finding disease genes and variants . . . 219
5.9 Key concepts . . . 223
5.10 Recommended reading . . . 224
5.11 Background reading (open access) . . . 224
5.12 Student recommendations . . . 226
5.13 Review problems . . . 226

6 Genomes **238**
6.1 Overview . . . 238
6.2 Background: genes, genomes and cells . . . 238
6.3 From the first genome projects . . . 240
6.4 Differences across the tree of life . . . 243
6.5 Gene order, synteny and genome rearrangements . . . 247

6.6 Genomes in populations . . . 251
6.7 Repeats and the 3D genome . . . 254
6.8 Pseudogenes, new genes and small proteins . . . 257
6.9 RNA genes and regulatory elements . . . 260
6.10 Key concepts . . . 262
6.11 Recommended reading . . . 263
6.12 Background reading (open access) . . . 264
6.13 Student recommendations . . . 265
6.14 Review problems . . . 265

7 Expression analysis **273**
7.1 Overview . . . 273
7.2 The first genome-wide methods: microarrays . . . 273
7.3 RNA sequencing . . . 275
7.4 Testing and statistics . . . 279
7.5 From gene lists to biology . . . 286
7.6 Single-cell genomics . . . 288
7.7 Single-cell sequencing and cancer evolution . . . 292
7.8 Key concepts . . . 294
7.9 Recommended reading . . . 295
7.10 Background reading (open access) . . . 296
7.11 Student recommendations . . . 297
7.12 Review problems . . . 297

8 Proteomics **306**
8.1 Overview . . . 306
8.2 Mass spectrometry methods . . . 306
8.3 Algorithms for peptide identification . . . 311
8.4 Statistics and error control . . . 314
8.5 Proteomics beyond protein abundance . . . 318
8.6 Proteomics without mass spectrometry . . . 320
8.7 Repositories and standards . . . 323
8.8 Key concepts . . . 323
8.9 Recommended reading . . . 324
8.10 Background reading (open access) . . . 325
8.11 Student recommendations . . . 326
8.12 Review problems . . . 326

9 Protein structure **334**
9.1 Overview . . . 334
9.2 CASP: a blind test of structure prediction . . . 334
9.3 Background: amino acids, energetics and enzymes . . . 336
9.4 Protein structure and how it is determined . . . 338
9.5 Protein folding . . . 341
9.6 Structure comparison and evolution . . . 344
9.7 Protein structure databases . . . 345
9.8 Protein structure prediction . . . 346
9.9 Protein language models and structure . . . 353
9.10 Membrane proteins . . . 355
9.11 Dynamics and docking . . . 356
9.12 Protein design . . . 360
9.13 Key concepts . . . 361

9.14 Recommended reading . . . 363
9.15 Background reading (open access) . . . 363
9.16 Student recommendations . . . 365
9.17 Review problems . . . 365

10 Protein function **374**
10.1 Overview . . . 374
10.2 Definitions of function . . . 374
10.3 Experimental determination of function . . . 377
10.4 Function prediction . . . 380
10.5 Subcellular localisation . . . 384
10.6 Predicting the effects of mutations . . . 386
10.7 Benchmarking function prediction . . . 388
10.8 Key concepts . . . 390
10.9 Recommended reading . . . 391
10.10 Background reading (open access) . . . 392
10.11 Student recommendations . . . 392
10.12 Review problems . . . 392

11 Protein–protein interactions **401**
11.1 Overview . . . 401
11.2 Experimental methods . . . 401
11.3 Networks . . . 406
11.4 Prediction methods . . . 412
11.5 Key concepts . . . 414
11.6 Recommended reading . . . 416
11.7 Background reading (open access) . . . 416
11.8 Student recommendations . . . 417
11.9 Review problems . . . 417

12 Machine learning **425**
12.1 Overview . . . 425
12.2 Basics . . . 425
12.3 A short history of machine learning in bioinformatics . . . 431
12.4 Deep learning . . . 433
12.5 Generative models . . . 437
12.6 Foundation models . . . 438
12.7 AI in research practice . . . 440
12.8 Integrating data . . . 441
12.9 Key concepts . . . 444
12.10 Recommended reading . . . 446
12.11 Background reading (open access) . . . 446
12.12 Student recommendations . . . 447
12.13 Review problems . . . 447

13 Image analysis and models of cells **456**
13.1 Overview . . . 456
13.2 Images as data . . . 456
13.3 Deep learning for images . . . 458
13.4 Image-based profiling . . . 459
13.5 Protein localisation from images . . . 461
13.6 Tissues: multiplexed imaging and computational pathology . . . 462

13.7 Cell atlases and virtual cells . . . 463
13.8 Whole-cell models and cryo-electron tomography . . . 464
13.9 Key concepts . . . 465
13.10 Recommended reading . . . 466
13.11 Background reading (open access) . . . 466
13.12 Student recommendations . . . 468
13.13 Review problems . . . 468

References **475**

Appendices **591**

A Recommended reading **591**
A.1 Sequence searching . . . 591
A.2 Databases . . . 619
A.3 Evolution . . . 680
A.4 Phylogeny . . . 708
A.5 Genomics . . . 731
A.6 Genomes . . . 743
A.7 Expression analysis . . . 757
A.8 Proteomics . . . 822
A.9 Protein structure . . . 860
A.10 Protein function . . . 923
A.11 Protein–protein interactions . . . 954
A.12 Machine learning . . . 979
A.13 Image analysis and models of cells . . . 1006

B Further papers **1105**
B.1 Sequence searching . . . 1105
B.2 Databases . . . 1123
B.3 Evolution . . . 1134
B.4 Phylogeny . . . 1154
B.5 Genomics . . . 1174
B.6 Genomes . . . 1195
B.7 Expression analysis . . . 1214
B.8 Proteomics . . . 1237
B.9 Protein structure . . . 1260
B.10 Protein function . . . 1290
B.11 Protein–protein interactions . . . 1305
B.12 Machine learning . . . 1329
B.13 Image analysis and models of cells . . . 1342

C Textbooks and general resources **1365**
C.1 Textbooks . . . 1365
C.2 Keeping up to date . . . 1365
C.3 Collections of papers . . . 1365

D Further explanations **1367**
D.1 Sequence searching . . . 1367
D.2 Databases . . . 1382
D.3 Evolution . . . 1404

D.4 Phylogeny . . . 1439
D.5 Genomics . . . 1459
D.6 Genomes . . . 1507
D.7 Expression analysis . . . 1522
D.8 Proteomics . . . 1580
D.9 Protein structure . . . 1619
D.10 Protein function . . . 1638
D.11 Protein–protein interactions . . . 1677
D.12 Machine learning . . . 1709
D.13 Image analysis and models of cells . . . 1752

E Wikipedia index **1784**
E.1 Sequence searching . . . 1784
E.2 Databases . . . 1788
E.3 Evolution . . . 1791
E.4 Phylogeny . . . 1794
E.5 Genomics . . . 1797
E.6 Genomes . . . 1800
E.7 Expression analysis . . . 1803
E.8 Proteomics . . . 1806
E.9 Protein structure . . . 1810
E.10 Protein function . . . 1815
E.11 Protein–protein interactions . . . 1817
E.12 Machine learning . . . 1819
E.13 Image analysis and models of cells . . . 1822

F Cross-topic problems and case studies **1826**
F.1 Cross-topic problems . . . 1826
F.2 Case study A: from sequence to mechanism . . . 1827
F.3 Case study B: from cohort to biological claim . . . 1828

About this guide

> ⚠ Disclaimer
>
> This book is **not complete**. It is a study aid, meant to help you organise your preparation for the exam. We do not guarantee that it covers everything you need to know to pass Stora Tentan, and the exam is not limited to what is written here. Use it as a map of the field and a starting point for your own reading, together with the papers, textbooks and lectures it points to, and discussions with other students and your supervisor.

> ❗ Version 0.6.1 (October 2026)
>
> This is an early version of the book. Version 0.6.1 lists the classical papers in bioinformatics under Appendix C. Version 0.6 filled gaps found by a comparison with two classic textbooks (Pevsner; Zvelebil & Baum): new sections on human genetics (linkage disequilibrium, GWAS, fine-mapping, polygenic scores, diagnosing rare disease), regulatory and epigenomic sequencing, whole-genome alignment and conservation scores, pseudogenes, new genes and small proteins, and the experimental determination of function, a short introduction to information theory, and new subsections in most chapters. Version 0.5 integrated Matteo Tadiello's companion textbook (review problems, figures and new text), and version 0.4 added a chapter on image analysis and models of cells (Chapter 13), which is not part of the 2025 exam description. The chapters and reading lists are open for comments from teachers and students: use the suggestion form (SU login).

This book collects reading material for **Stora Tentan**, the 15 hp oral exam in bioinformatics for PhD students at the Department of Biochemistry and Biophysics (DBB), Stockholm University. It is built on the current exam description (2025) and on the more detailed reading lists that were used in earlier years. It is not a course book: everything in it is chosen for one purpose, to help PhD students at DBB prepare for this exam.

What the exam is

Stora Tentan should show that a student awarded a PhD from DBB has a good basic knowledge of bioinformatics, *wider than their own thesis area.* The exam assumes the knowledge needed to pass the department's advanced-level courses in bioinformatics and comparative genomics, but it is primarily a test of the ability to explain how bioinformatics methods are used to answer fundamental biological questions. Students are also expected to follow seminars and dissertations at DBB and SciLifeLab and to be aware of current international research.

- Two examiners, about two hours.
- No single required textbook. Zvelebil & Baum (Zvelebil and Baum 2008) or Pevsner (Pevsner 2015) cover most areas, but they are not up to date — you need to update your knowledge to the current state of the art (the proceedings of the last few ISMB and ECCB meetings are a good start).

- The exam is **not** about knowing every detail. You should be able to discuss and reason about methods: how they work (algorithms), their advantages and limitations, and when to use them.
- Fundamental knowledge of the algorithms and limitations of established methods for **sequence searching, protein structure prediction, phylogeny and function annotation** is required.
- Every exam is unique. Depending on how the first topic goes, other topics may be covered in more or less depth, or not at all.

Experience shows that discussing the exam with other PhD students is a very efficient way to prepare, and that online lectures often explain things better than textbooks.

Programming and reproducible research

Programming and reproducibility are **general professional background**, not a separate exam topic. You will not be asked to write code, and the topic chapters do not return to these practices in detail. But you are expected to know how computational research is done and to be able to discuss the tools and practices you use: examiners may ask how you would implement or scale an algorithm, or how someone else could reproduce one of your results.

What you should know

- **Languages and their roles.** Python and R for analysis, statistics and visualisation; the Unix shell for handling files and running programs; compiled languages (C, C++, Rust) for performance-critical tools such as aligners. Know why most heavy bioinformatics tools are written in compiled languages, while analyses are written in scripting languages.
- **Algorithms and complexity.** Big-O notation, and why an $O(n^2)$ algorithm becomes unusable on a million sequences while an $O(n \log n)$ one does not. Basic data structures (arrays, hash tables, trees, graphs) and the algorithmic ideas that recur in this book: dynamic programming (Section 1.3), hashing and indexing (Section 1.4), and graph algorithms (Chapter 5).
- **Computing environments.** Working on a Unix system and on a computer cluster with a job scheduler such as Slurm; when to use GPUs (mainly for deep learning); cloud computing and its cost.
- **Version control.** Using git for code and analyses, and platforms such as GitHub or GitLab for collaboration (Perez-Riverol et al. 2016).
- **Reproducibility.** Recording software versions and parameters; managing software environments (conda, containers such as Docker or Apptainer); using workflow managers such as Snakemake or Nextflow (Köster and Rahmann 2012; Di Tommaso et al. 2017); documenting analyses, for example in notebooks or Quarto/R Markdown documents. Sandve et al. (2013) and Wilson et al. (2017) give short, practical rules.
- **Reproducible is not the same as correct.** An analysis is *reproducible* if the same data and code give the same result, *replicable* if new data give the same conclusion, and *robust* if the conclusion survives other reasonable analysis choices. A perfectly reproducible analysis can still be wrong, for example because of data leakage or confounding (4). Containers record the software, but not the input data, the reference files or sources of randomness, which must be recorded separately.
- **Data management.** The FAIR principles (findable, accessible, interoperable, reusable) (Wilkinson et al. 2016); depositing data in public repositories (Chapter 2); standard file formats (FASTA, FASTQ, SAM/BAM, VCF, PDB/mmCIF; see Table 1); handling sensitive human data in line with GDPR and ethical approvals.
- **Testing and correctness.** Checking code on small examples with known answers, and being sceptical of results that look too good. Many published errors come from simple bugs such as off-by-one errors, swapped sample labels or spreadsheet conversions of gene names to dates.

Common file formats

Most bioinformatics data are exchanged in a small number of community formats. You should know what each contains and what it is used for.

Table 1: Common bioinformatics file formats.

Format	What it contains
FASTA	Sequences (nucleotide or protein), each with a header line starting with >; no quality information
FASTQ	Sequencing reads with a quality score for every base (Phred scale, encoded as characters), four lines per read
SAM / BAM / CRAM	Reads aligned to a reference: position, mapping quality, flags and a CIGAR string describing the alignment. SAM is text, BAM its compressed binary form, CRAM a smaller, reference-based compression
VCF (and binary BCF)	Variants: position, reference and alternative alleles, quality, filters, and the genotype of each sample
BED	Genomic intervals (chromosome, start, end, and optionally name, score and strand), for example peaks or target regions
GFF3 / GTF	Genome annotation: genes, transcripts, exons and coding sequences, with parent–child relations in an attribute column. GTF is an older variant used by Ensembl and GENCODE
PDB / mmCIF	Atomic coordinates of macromolecular structures. The legacy fixed-column PDB format cannot hold very large structures; PDBx/mmCIF is the archive standard
HDF5, AnnData (`.h5ad`)	HDF5 is a hierarchical binary container for large arrays. AnnData, built on it, stores a single-cell count matrix together with cell and gene annotations and embeddings
OME-Zarr	Chunked, cloud-friendly storage of large multidimensional images, for example microscopy and spatial omics, with standard metadata

Coordinates differ between formats: BED (and BAM internally) counts from 0 with half-open intervals, whereas SAM, VCF and GFF count from 1 with closed intervals, so conversions are a classic source of off-by-one errors (Chapter 2).

A provenance checklist

A computational result can only be reproduced if everything that produced it is recorded. For any analysis, you should be able to state:

- the **data**: accession numbers, and the **release or date of every database** used (a UniProt or Pfam release changes search results; see Chapter 2);
- the **reference genome and annotation versions**, for example GRCh38 with a specific GENCODE release, or the T2T or pangenome reference (Chapter 5);
- the **software versions and all parameters**, including thresholds and defaults that were left unchanged;
- **random seeds**, for anything stochastic: MCMC, clustering, UMAP, training of machine-learning models, structure prediction with several seeds;
- the **computing environment**: a conda environment file or a container image;

- the **models**: the versions or weights of trained models, and for AI tools the tool, model version, date and prompts (Section 12.7);
- the **workflow**: the scripts or workflow-manager pipeline that link these steps, under version control;
- where the **code and data are deposited**, so that others can rerun the analysis.

Reading

- Sandve et al. (2013) — ten simple rules for reproducible computational research.
- Wilson et al. (2017) — good enough practices in scientific computing.
- Software Carpentry lessons, for example the Unix shell, version control with git, and programming in Python or R.

Uncertainty: many numbers, different meanings

Almost every method in this book reports a number that expresses confidence in a result. They answer different questions and must not be confused, which is a frequent source of mistakes in oral exams. The table summarises the most important ones.

Table 2: Measures of confidence used in this book.

Measure	What it is	Common misreading	See
E-value	Expected number of hits with at least this score by chance, for one query against this database	A probability; it grows with database size	Section 1.4
p-value	Probability of data at least as extreme as observed, *if* the null hypothesis is true	The probability that the null hypothesis is true	Section 7.4.1
FDR	Expected *proportion* of false positives among the set of results called significant	The probability that one particular hit is false	Section 7.4
q-value	For one result, the smallest FDR threshold at which it would be called significant	The FDR of that single result	Section 7.4
Local FDR, posterior error probability	Probability that one particular result is false, under the model	Model-free; it is only as good as the model	Section 8.4
Bootstrap support	Percentage of resampled data sets in which a clade is recovered	The probability that the clade is true	Section 4.3.9
Posterior probability (Bayesian)	Probability of a clade or parameter given the data, the model and the priors	Independent of the model; it can be close to 1 for wrong clades if the model is wrong	Section 4.3.9
pLDDT, PAE, ipTM	AlphaFold's predicted error of local structure, relative positions and interfaces	Measured accuracy; confident predictions can still be wrong	Section 9.8
Classifier score	A model's output for one example	A probability: it is one only if the model is calibrated, and it depends on how common positives are	4

All of them are conditional on assumptions: the null model, the substitution model, the training data. A small number is not evidence by itself. It is evidence only together with the assumptions behind it.

Information theory in brief

Several quantities in this book come from information theory, founded by Shannon (Shannon 1948). With logarithms to base 2 they are measured in **bits**; with natural logarithms, as in most machine learning, in *nats*.

Entropy measures the uncertainty of a distribution p over symbols a, or the average surprise $-\log_2 p_a$ of an observation:

$$H(p) = -\sum_a p_a \log_2 p_a.$$

It is the average number of yes/no questions needed to identify a symbol. A uniform distribution over the four nucleotides has 2 bits, over the 20 amino acids $\log_2 20 \approx 4.3$ bits, and a fully conserved alignment column has 0 bits. A related quantity, the information content $-\log_2 P(t)$ of a single event, weights rare, specific GO terms more than general ones in the evaluation of function prediction (Section 10.6).

Relative entropy, or the **Kullback–Leibler divergence** (Kullback and Leibler 1951), measures how different a distribution p is from a reference q:

$$D(p \,\|\, q) = \sum_a p_a \log_2 \frac{p_a}{q_a}.$$

It is the average log-odds score $\log_2(p_a/q_a)$ when the data really come from p. It is never negative, and zero only if $p = q$, but it is not symmetric, so it is not a true distance. It appears twice in Chapter 1. The relative entropy of a substitution matrix is the average score per aligned pair in true alignments, about 0.7 bits for BLOSUM62, and tells how long an alignment must be to stand out from chance (Section 1.3). The information content of a motif position is the relative entropy of the column with respect to the background, and is the height of a stack in a sequence logo (Section 1.6).

Mutual information measures how much knowing one variable reduces the uncertainty about another. It is the relative entropy between the joint distribution and the product of the marginals:

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)}.$$

It is zero only if X and Y are independent, and unlike correlation it also detects non-linear dependence. Between two columns of an alignment it reveals covariation: base pairs in RNA (Section 1.7) and residue contacts in proteins (Section 3.5, Section 9.8), where global models are needed to separate direct from indirect couplings. It is also used to infer gene regulatory networks from expression data (Section 11.2). The relative entropy of a substitution matrix is itself the mutual information between the two residues of a true aligned pair.

Cross-entropy compares the true distribution p with a model's predicted distribution q:

$$H(p, q) = -\sum_a p_a \log q_a = H(p) + D(p \,\|\, q).$$

For a classifier, p puts all its weight on the true class y, and the loss for one example is $-\log q_y$, the negative log-likelihood of the correct answer. Minimising the average cross-entropy loss is therefore maximum likelihood, and since $H(p)$ is fixed by the data, it also minimises the divergence of the model from the data (4).

Perplexity is the exponential of the cross-entropy, 2^H in bits or e^H in nats. It is the effective number of equally likely choices the model is choosing between at each position: a uniform guess over the amino acids gives 20, and a perfect prediction gives 1. It is the usual measure of how well a protein language model predicts sequences (Section 1.10).

> **Sources**
>
> Written for this book. For background, see the Wikipedia articles Entropy (information theory), Kullback–Leibler divergence, Mutual information, Cross-entropy and Perplexity.

How each chapter is organised

The thirteen topic chapters share the same structure:

1. **Overview** — what the chapter covers and how it connects to the others.
2. **Sections** — a full text on each part of the topic, written mainly from Wikipedia and other open sources (credited at the end of each section), with references to the key papers. Background that every student is expected to know, such as statistics, basic biochemistry and the use of AI tools, is included where it is needed: statistics in Chapter 7, nucleic acids and cells in Chapter 6, amino acids, energetics and enzymes in Chapter 9, and AI in research practice in Chapter 12.
3. **Key concepts** — a checklist of things you should be able to explain.
4. **Recommended reading** — about six papers to read in full, mostly from the reading lists of earlier years, each with a one-line note on why it matters.
5. **Background reading (open access)** — overviews to get started, mostly Wikipedia articles and tutorials, each with a short description.
6. **Student recommendations** — papers, online lectures, tutorials and exam experiences recommended by students. Students can send suggestions, including corrections and exam questions, through the suggestion form (SU login); the teachers review them and add the good ones in the next version.
7. **Review problems** — problems that test understanding rather than recall, each with extensions, a solution outline and common errors (click to show them in the HTML version). They are adapted from the companion textbook by Matteo Tadiello (see below).

The papers are collected in appendices, so that the chapters stay readable:

- **Appendix A, Recommended reading** (A.1–A.13, one section per chapter): the recommended papers with titles, abstracts, DOI and library links. In the PDF version the open-access papers follow in full.
- **Appendix B, Further papers** (B.1–B.13): for each chapter, a longer list of papers grouped by theme, with titles, DOI links and abstracts (click *Abstract* in the HTML version). Papers marked ⋆ were mandatory in earlier reading lists.
- **Appendix D, Further explanations** (D.1–D.13): for each section of a chapter, one to three open-access reviews, tutorials or textbook chapters that explain the topic in more detail than the book does, each with a short comment. In the PDF version one or two of them per chapter follow in full. Each section of the chapters points to its part of appendix D.
- **Appendix E, Wikipedia index** (E.1–E.13): the main Wikipedia articles for each section, with a one-line description. They are a quick way to look up a term, not a source to cite.
- **Appendix F, Cross-topic problems and case studies**: problems that combine several chapters, as questions in the oral exam often do, and two longer case studies.

Appendix C lists textbooks and general resources. Full references are collected in References; hover over a citation to see it.

Topics

Topic	What is covered
Sequence searching	Alignment algorithms, substitution matrices, BLAST, profile HMMs, significance, MSA, motifs
Databases	Primary and derived databases, their hierarchy and relations
Evolution	Homology, orthology/paralogy and orthology databases, gene duplication, selection, drift
Phylogeny	Distance, parsimony, likelihood and Bayesian methods; models; support
Genomics	Sequencing technologies, assembly, read mapping, variant calling, gene prediction
Genomes	Prokaryotic vs. eukaryotic genome organisation, viral genomes, gene order and synteny, genome size and complexity
Expression analysis	Microarrays, RNA-seq, single-cell, normalisation, multiple testing
Proteomics	Mass spectrometry, peptide identification, quantification, FDR
Protein structure	Structure, domains, secondary/tertiary prediction, AlphaFold, design, docking
Protein function	Function definitions (GO), annotation transfer, localisation, variant effects
Protein–protein interactions	Experimental and computational PPI methods, networks, pathway and network enrichment
Machine learning	HMMs, neural networks, evaluation and overfitting, deep learning, language models, integrating data
Image analysis and models of cells	Bioimage analysis, segmentation and classification, Cell Painting, protein localisation, pathology, cell atlases, virtual and whole-cell models (not in the 2025 exam description)

Earlier versions of the instructions

The reading lists from earlier years (in `data/`) contained a list of mandatory articles and a much longer list of background papers. These have been distributed over the topic chapters. A 2021 variant of the exam was instead based on Pevsner's textbook (Pevsner 2015) together with the review by Sapoval et al. (2022); both remain excellent starting points.

Additional material

A lot of the material in this book is covered by the two bioinformatics courses given by Arne Elofsson (KB8029) and Erik Sonnhammer (KB8030). The course book for KB8029 can be found at https://course.bioinfo.se. The suggested course litterature for KB8030 is *Zvelebil, M. J., Baum, J.O. (2007) Understanding Bioinformatics, Garland Science*

Companion textbook and contributors

Matteo Tadiello, a student of Arne Elofsson, has written a companion textbook, *Modern Bioinformatics: From Foundations to Frontiers* (<https://bioinfo.se/teaching/modern-bioinformatics/>). It covers the same topics as a general textbook, in more depth and with more mathematics, and is a good next step after the summaries in this book. Each chapter here links to the matching chapters of his book. The review problems at the end of each chapter, the cross-topic problems and case studies in appendix F, many of the figures, and new text in most chapters are adapted from his book with his permission (CC BY 4.0). This includes the sections on orthology databases, gene order and synteny, viral genomes, pathway and network enrichment, and integrating data; every adapted part is credited where it appears.

Licence

This book is licensed under the Creative Commons Attribution-ShareAlike 4.0 International licence (CC BY-SA 4.0). Parts of the text are adapted from Wikipedia and other openly licensed sources, which are credited in each section. Included papers keep their own licences, and only openly licensed papers are reproduced in full. A private edition of the book (rendered with `quarto render --profile private`) also contains copyrighted papers; it is for students at DBB only and must not be distributed.

Part I

Topics

1 Sequence searching

> **From the exam description**
>
> Both historical and state-of-the-art methods are used for speed and accuracy. Understanding the advantages of different algorithms is essential.

1.1 Overview

Finding sequences that are similar to a query is one of the most frequently used operations in bioinformatics. It is the first step in annotating a new gene, building a family alignment, inferring a phylogeny or predicting a structure. This chapter follows the development of the field:

- what homology is, and how it relates to sequence similarity (1);
- exact alignment by dynamic programming and the substitution matrices that score it (Section 1.3);
- the heuristic and indexing methods that make database search fast, and the statistics that tell us which hits to trust (Section 1.4);
- comparing sequences without aligning them, by their k-mers (Section 1.5);
- multiple alignments, motifs, profiles and profile hidden Markov models, which capture whole families and find more distant relationships (Section 1.6);
- how the same ideas are extended to structured RNAs, where secondary structure is often conserved better than sequence (Section 1.7);
- aligning whole genomes, and the conservation scores computed from multiple genome alignments (Section 1.8);
- protein domains and the databases that classify them (Section 1.9);
- protein language models and how they are starting to be used to detect remote homology (Section 1.10);
- how search methods are benchmarked, and what can go wrong (Section 1.11).

1.2 Homology and sequence similarity

Almost everything in this chapter rests on one idea. If two genes or proteins descend from a common ancestor, then what we know about one tells us something about the other. The relationship is called *homology*, and sequence searching is, at bottom, a set of methods for inferring homology from sequence similarity. Before turning to the algorithms, it is worth being precise about what is inferred, from what, and why the inference works.

1.2.1 Evolutionary background

All living organisms descend from a common ancestor, and their genomes have been shaped by the same small set of processes (see Chapter 3 for a fuller treatment):

- **Mutation** creates variation. The simplest mutations are *point substitutions*, the replacement of one nucleotide by another. In DNA these are classified as *transitions* (purine ↔ purine or pyrimidine ↔ pyrimidine), which are more frequent, and *transversions* (purine ↔ pyrimidine). *Insertions and deletions* (indels) add or remove nucleotides. Larger events duplicate, delete, invert or move whole segments of DNA, from a few nucleotides up to entire genomes.
- **Recombination** reshuffles existing variation. In proteins, unequal crossing-over and exon shuffling can join segments that encode separate domains, so that new proteins arise as new combinations of old parts (Section 1.9).
- **Horizontal (lateral) gene transfer** moves genes between organisms other than by descent. It is common in prokaryotes and rarer, but not absent, in eukaryotes.
- **Natural selection** and **genetic drift** determine the fate of each variant. Most mutations that change the structure or function of an important protein are harmful and are removed by *purifying selection*. Mutations with little effect are fixed or lost largely at random by drift, as the neutral theory proposed (Kimura 1968). Occasionally a mutation is beneficial and spreads by *positive selection*.

For sequence analysis, the key consequence is that sequences diverge gradually and unevenly. Positions that matter for structure or function (the hydrophobic core, catalytic residues, binding sites) change slowly. Surface loops and positions that tolerate many amino acids change quickly. When substitutions do occur at constrained positions, they tend to be *conservative*, replacing an amino acid by one with similar properties (isoleucine by leucine or valine, aspartate by glutamate). These patterns are exactly what substitution matrices capture (Section 1.3) and what profiles capture position by position (Section 1.6).

Because divergence is gradual, the similarity between two homologous sequences decays with the time since they shared an ancestor. Eventually so many substitutions have occurred, often several at the same site, that the remaining similarity is no longer distinguishable from chance. The limits of sequence searching are set by this decay.

1.2.2 Genes and proteins

In classical genetics, a *gene* is a unit of heredity, a factor that determines a trait. The molecular definition, used in this book, is a segment of DNA that is transcribed to produce a functional RNA. Most of the genes analysed in this chapter are *protein-coding*: the RNA is a messenger RNA (mRNA) that is translated into a protein. *Non-coding genes* produce functional RNAs such as ribosomal and transfer RNAs, spliceosomal RNAs and microRNAs. The flow of information from DNA to RNA to protein is the *central dogma* of molecular biology.

In eukaryotes a protein-coding gene is usually split. Coding *exons* alternate with non-coding *introns*, which are removed by splicing. Alternative splicing allows one gene to produce several protein *isoforms*. In prokaryotes genes are mostly continuous and are often organised in operons (see Chapter 6).

A *protein* is one or more chains of amino acid residues joined by peptide bonds. The sequence of the chain is specified by the gene through the genetic code. Each of the 64 codons (nucleotide triplets) codes for one of 20 standard amino acids or for a stop signal. Two non-standard amino acids, selenocysteine and pyrrolysine, are used in some organisms. The code is *degenerate*: most amino acids are encoded by several codons that usually differ in the third position. As a result, many DNA substitutions are *synonymous* and leave the protein unchanged. The amino-acid sequence determines how the chain folds into a three-dimensional structure, and the structure determines the function (Anfinsen 1973). Proteins are typically 100–1,000 residues long and many consist of several *domains*: compact units that fold, function and evolve semi-independently (Section 1.9). Short chains of fewer than 20–30 residues are usually called peptides.

Two properties of proteins matter particularly for sequence comparison. First, the 20-letter alphabet carries much more information per position than the 4-letter nucleotide alphabet. Second, among homologous proteins the three-dimensional structure tends to be conserved much longer during evolution than the sequence that encodes it (an empirical tendency, not a law). We return to both points below.

1.2.3 Homology: a statement about ancestry

The word *homology* comes from the Greek *homologos*, "agreeing". It was introduced into biology by the anatomist Richard Owen in 1843. He used it for "the same organ in different animals under every variety of form and function", and contrasted it with *analogy*, different structures that perform the same function. Owen's definition was not evolutionary. In 1859 Darwin explained homologous structures as inheritance from a common ancestor, and that is how the term has been used ever since. The classical example is the vertebrate forelimb. The wing of a bat, the flipper of a whale, the foreleg of a horse and the human arm are built from the same bones (humerus, radius, ulna) inherited from an ancestral tetrapod limb, even though they now serve very different purposes.

The molecular definition is the same. Two genes, proteins or parts of them are **homologous if they descend from a common ancestral sequence**. Homology is therefore a statement about history, not about the sequences as we observe them today. It has three consequences that are often misunderstood:

1. **Homology is qualitative.** Two sequences either share an ancestor or they do not. There is no such thing as "45% homology". The quantity that can be measured is *similarity*, for example the percentage of identical residues in an alignment (*percent identity*) or the percentage of similar residues (*percent similarity*). Similarity is the observation; homology is the conclusion drawn from it.
2. **There is only one kind of homology.** Expressions such as "structural homology" or "functional homology" mix up the evidence with the conclusion. Two proteins can be *structurally similar* because they are homologous, or because the same fold evolved twice. They can have the same function for the same two reasons. Structural similarity, like sequence similarity, is *evidence* for homology, sometimes the strongest evidence available. But it is not a separate type of homology.
3. **Homology need not cover the whole sequence.** Many proteins are mosaics of domains with different histories. Two proteins can be homologous over one domain and unrelated elsewhere, a situation sometimes called *partial homology*. Statements about homology should therefore refer to aligned regions or domains, not automatically to whole proteins.

These points were made forcefully by Reeck et al. (1987). Their paper, " 'Homology' in proteins and nucleic acids: a terminology muddle and a way out of it", was written by eleven leading molecular biologists and evolutionists, among them Doolittle, Fitch, Margoliash, Jukes and Zuckerkandl. It noted that a large fraction of papers used "homology" to mean similarity, and argued that the word should be reserved for common ancestry. The misuse persists nearly 40 years later, and you should expect to be asked about it.

> Example: how to report a result
>
> *Wrong:* "Protein X is 38% homologous to protein Y."
> *Right:* "Protein X and protein Y share 38% sequence identity over 240 aligned residues (E-value 10^{-35}). The similarity is far higher than expected by chance, so we conclude that they are homologous."

1.2.4 Orthologues, paralogues and xenologues

Homologous genes are classified according to the type of event at the node where their histories split. The terms were introduced by Walter Fitch (Fitch 1970) and later refined (Fitch 2000; Koonin 2005):

- **Orthologues** are homologues that diverged through a *speciation* event. When an ancestral species splits into two, the copies of a gene in the two daughter species are orthologues. Human and mouse β-globin (*HBB* and *Hbb-b1*) are orthologues. So are the cytochrome *c* genes of humans and yeast.
- **Paralogues** are homologues that diverged through a *gene duplication*. Human α-globin and β-globin are paralogues, produced by a duplication that occurred roughly 450 million years ago in an early vertebrate. Myoglobin is an even older paralogue of both. Paralogues are present in the same genome, but because the duplication can precede later speciations, a gene in one species can also be paralogous to a gene in another species (human α-globin and mouse β-globin, for example).
- **Xenologues** are homologues whose history includes a *horizontal gene transfer*. Antibiotic-resistance genes carried on plasmids, which have spread between distantly related bacteria, are typical examples.

Paralogy is further divided relative to a particular speciation event (Remm et al. 2001):

- **In-paralogues** (symparalogues) arose by duplication *after* the speciation, so they are all co-orthologous to the corresponding gene or genes in the other species.
- **Out-paralogues** (alloparalogues) arose by duplication *before* it.

For example, if a gene duplicated in the primate lineage after the split from rodents, the two human copies are in-paralogues relative to the human–mouse comparison. Both are orthologous to the single mouse gene. Paralogues produced by a *whole-genome duplication* are called **ohnologues**, after Susumu Ohno, who proposed that two rounds of genome duplication early in vertebrate evolution ("2R") supplied the raw material for vertebrate complexity (Ohno et al. 1968).

The distinction matters in practice because of how function evolves. After a speciation, orthologues usually continue to perform the same role in their respective species. After a duplication, the selective constraints on one or both copies may be relaxed. A copy is often lost, but it may also acquire a new function (*neofunctionalisation*), or the two copies may divide the ancestral functions between them (*subfunctionalisation*) (Lynch and Conery 2000). The resulting rule of thumb, that orthologues are more likely than paralogues to share function (sometimes called the *ortholog conjecture*), underlies the transfer of functional annotation between species (Chapter 10). It is a useful generalisation, not a law, and has been debated in the literature. The classic counterexample among paralogues is angiogenin, which derives from the pancreatic ribonuclease family but has a different biological function.

Orthology and paralogy are properties of a *gene tree reconciled with a species tree*. In principle they are determined by phylogenetic analysis (Chapter 4). In practice they are often inferred at genome scale with faster methods based on pairwise similarity. Examples are reciprocal best hits, COGs, InParanoid and OrthoFinder (Tatusov et al. 1997; Remm et al. 2001; Emms and Kelly 2019; Altenhoff et al. 2019). These methods are discussed in Chapter 3. Note that *similarity alone cannot tell orthologues from paralogues*. The most similar sequence in another genome is often, but not always, the orthologue, for instance when the true orthologue has been lost.

1.2.5 Analogy and convergent evolution

Similar features can also evolve independently. Such features are **analogous**, and the process is called **convergent evolution** (or *homoplasy* in cladistics). Insect and bird wings are analogous: both support

powered flight but evolved separately. The same feature can be homologous at one level and analogous at another. Bird and bat wings are analogous *as wings*, since the common ancestor did not fly, but homologous *as forelimbs*.

At the molecular level convergence is most convincing for *functional features*, especially active sites, because chemistry constrains the possible solutions. The textbook example is the **serine proteases**. Chymotrypsin (and trypsin, elastase and thrombin) and the bacterial protease subtilisin use the same Ser–His–Asp *catalytic triad*, arranged almost identically in space. Yet the two families have completely different folds: chymotrypsin is built from two β-barrels and subtilisin is an α/β protein. The order of the triad residues along the chain differs too: His–Asp–Ser in chymotrypsin and Asp–His–Ser in subtilisin. The similarity is analogous. Catalytic triads have evolved independently in more than 20 enzyme superfamilies. Another example is carbonic anhydrase, where at least three structurally unrelated classes (α, β, γ) catalyse the same reaction.

Convergence at the level of a *full sequence* is, by contrast, extremely unlikely. There are 20^{100} possible sequences of 100 residues, and two unrelated proteins with a similar function have no reason to arrive at similar sequences. Convergence of *simple folds* (helical bundles, β-hairpins, perhaps some common α/β architectures) is more plausible, which is one reason why structural similarity is somewhat weaker evidence for homology than significant sequence similarity. Deciding whether two proteins with similar folds but no detectable sequence similarity are distant homologues or analogues is one of the hard problems in protein classification (Section 1.9). Arguments then rest on shared unusual features: rare structural details, conserved functional residues in equivalent positions, or the same arrangement of secondary-structure elements in an unusual order.

1.2.6 From similarity to homology

How, then, do we conclude that two sequences are homologous? The argument is statistical. When two sequences are aligned, some identical or similar residues will always be found by chance. The question is whether the similarity is *greater than expected for unrelated sequences* of the same length and composition. If it is substantially greater, the most parsimonious explanation is common ancestry, because convergence of whole sequences is so unlikely. Doolittle made this argument explicit in "Similar amino acid sequences: chance or common ancestry?" (Doolittle 1981). Modern search programs implement it by reporting, for each alignment, how often a score at least that high would be expected by chance in a database search (the E-value; see Section 1.4).

Doolittle was also one of the first to show how powerful the approach could be. In 1983 he and his colleagues compared the newly determined sequence of platelet-derived growth factor with their small database of known sequences. They found that it was almost identical to the product of *v-sis*, an oncogene of simian sarcoma virus. A growth factor and a cancer-causing gene turned out to be the same protein: a discovery made entirely by sequence comparison (Doolittle et al. 1983; Doolittle 2010).

The inference is **asymmetric**:

- *Significant similarity implies homology.* If the alignment score is far above what is expected by chance, the sequences are homologous, with a few well-known exceptions discussed below.
- *Lack of significant similarity does not imply non-homology.* Homologous sequences can diverge until their similarity is indistinguishable from noise. Many protein families that are clearly related by structure and function share no detectable sequence similarity.

A consequence is that the *sensitivity* of a search method, its ability to detect distant homologues, is a central measure of its quality. Much of the progress described in this chapter, from pairwise alignment through profiles and HMMs to structure- and language-model-based search, consists of pushing the detection limit further into the region where similarity is faint.

1.2.6.1 The twilight zone

How much identity is needed? Because unrelated sequences can be aligned with gaps, two random protein sequences typically show 15–20% identity over an optimal alignment, and short alignments can show much more. Any fixed threshold must therefore depend on the alignment length. Doolittle coined the term *twilight zone* for the range of pairwise identity in which homology can be neither confirmed nor ruled out by identity alone. The often-quoted range of roughly 20–30% applies only to alignments of typical domain length (about 100 residues or more); for short alignments the zone lies at much higher identity. The term is also used more loosely for the general difficulty of detecting distant homologues. Rost (1999) later quantified it on a large set of structurally verified pairs. He found that above a length-dependent curve, almost all pairs were homologous (same structure), while below it false positives rose sharply. For alignments of about 100 residues the threshold is near 30% identity, and for 250 residues or more it approaches 20% (Figure 1.1). Below about 20% lies what Rost called the *midnight zone* (Rost 1997). Here many pairs are still homologous, as their structures show, but percent identity carries essentially no information.

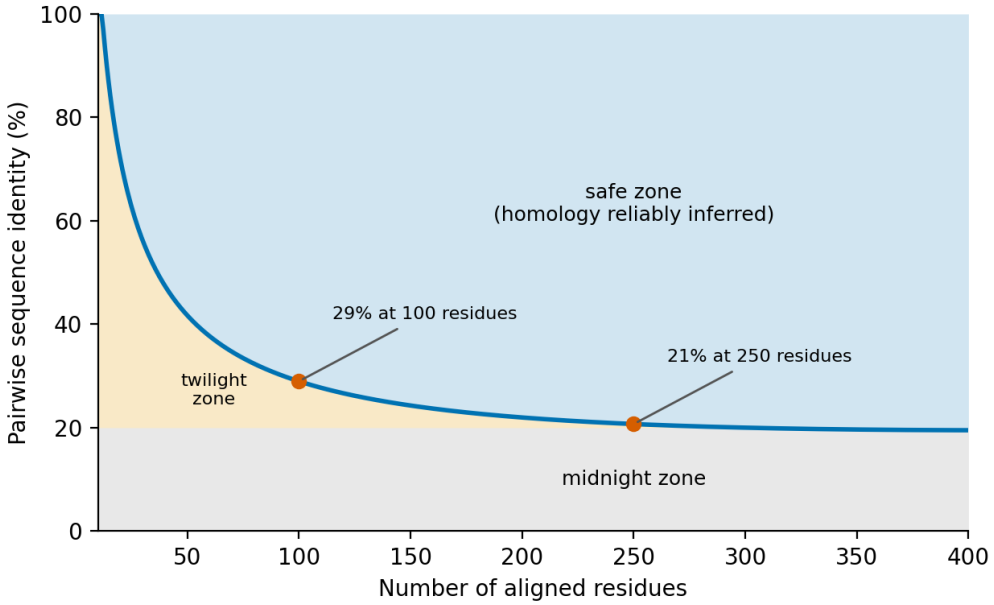


Figure 1.1: The twilight zone. The curve shows the length-dependent identity threshold of Rost (1999), $p(L) = 480\,L^{-0.32\,(1+e^{-L/1000})}$. Above it, homology can be inferred reliably from pairwise identity. Between the curve and about 20% identity (twilight zone), true and false relationships are mixed. Below about 20% (midnight zone), identity is uninformative even though many such pairs are homologous.

Percent identity is nevertheless a poor statistic for detecting homology. It ignores the type of substitution, is sensitive to how gaps are counted, and has no direct probabilistic interpretation. Scores computed with substitution matrices and gap penalties, converted to E-values, are much better. They are the basis of every method in this chapter. Percent identity remains useful for a different purpose: describing how *close* two known homologues are, for example when choosing a template for homology modelling (Chapter 9) or deciding whether annotation can be transferred.

1.2.6.2 Why compare proteins rather than DNA?

When a protein-coding gene is studied, it is almost always better to compare the translated protein sequences than the DNA. There are several reasons:

- **Alphabet size.** With 20 amino acids, a random match at one position occurs with a probability of about 6% (given typical amino-acid frequencies). With four nucleotides, it occurs about 25% of the time. The protein alphabet therefore carries more information per aligned position, and fewer positions are needed to reach significance.
- **Synonymous substitutions.** Because of the degeneracy of the genetic code, many nucleotide changes, especially at third codon positions, do not alter the protein. They accumulate quickly and erase DNA similarity even where the protein is conserved.
- **Substitution matrices.** Amino acids differ in their physicochemical properties, and the frequency with which one replaces another can be estimated from real alignments (Section 1.3). Conservative replacements can be scored as positive evidence. For DNA there is much less to exploit.
- **Composition and repeats.** Genomic DNA is full of repeats and low-complexity regions that produce spurious matches.

The effect is dramatic. DNA–DNA comparisons rarely detect homology between sequences that diverged more than a few hundred million years ago, whereas protein–protein comparisons routinely detect homologues that diverged more than two billion years ago, for example between bacteria and humans (Pearson 2013). DNA comparison is the right choice for non-coding sequences (RNA genes, regulatory regions), for very closely related sequences (resequencing, read mapping; see Chapter 5), and when the reading frame is unknown. Programs such as BLASTX and TBLASTN bridge the gap by translating nucleotide sequences in all six reading frames (Section 1.4).

1.2.6.3 Structure is more conserved than sequence

The three-dimensional structure of a protein changes much more slowly than its sequence. Pairs of homologous proteins with less than 20% sequence identity often still share the same fold, with most of their core superimposable within 2–3 Å (Chothia and Lesk 1986). Below about 30% identity, however, sequence similarity no longer reliably reveals the relationship. This is why structural comparison can detect homology far beyond the limits of sequence comparison. It is also why the structural classifications SCOP, CATH and ECOD group domains into superfamilies on the basis of structural and functional evidence that sequence alone cannot provide (Murzin et al. 1995; Holm and Sander 1996; Cheng et al. 2014). With accurate predicted structures now available for almost every known protein (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021; Varadi et al. 2021), structure-based search has become a practical complement to sequence search (Section 1.4, Chapter 9). The caveat from above still applies: a shared fold is strong but not conclusive evidence of common ancestry.

1.2.6.4 Pitfalls: similarity that is not homology

The rule "significant similarity implies homology" fails in a few well-characterised situations. All of them involve sequences that violate the assumptions behind the statistics, which assume sequences composed roughly at random from standard background frequencies:

- **Low-complexity regions**, such as stretches rich in a few amino acids (poly-Q, proline-rich or serine-rich segments) or short tandem repeats. These can give high scores between unrelated proteins. Search programs therefore mask them, for example with the SEG filter (Wootton and Federhen 1993), or adjust the scores for composition (Schäffer et al. 2001).
- **Biased composition and repetitive structural motifs**, such as coiled coils (heptad repeats), transmembrane helices, collagen-like repeats or signal peptides. These can produce significant-looking alignments between proteins that are related only by shared physical constraints.

- **Multidomain proteins.** A significant hit establishes homology only for the aligned region. Annotating a whole protein from a hit to one of its domains is a frequent source of error. So is assuming transitivity: if A is homologous to B over one domain and B to C over another, A and C need not be related at all. Before transferring an annotation, therefore, check *coverage* as well as significance: the fraction of the query (*query coverage*) and of the hit (*target* or *subject coverage*) that is included in the alignment. A highly significant hit that covers only one domain of a multidomain query supports transfer of that domain's properties, not of the function of the whole protein. Many annotation pipelines and clustering tools (e.g. MMseqs2) require a minimum coverage of both sequences for this reason.

In addition, an E-value always refers to a *particular search*: one query, a database of a given size and a scoring scheme. Two different issues should be kept apart. First, the E-value of a single query already includes the size of the database (the search space): the same alignment score gives a larger E-value in a larger database. $E = 0.01$ means that one chance hit this good is expected in a hundred such searches, whatever the database size. Second, when thousands of *independent queries* are searched, about 0.01 chance hits per query at this threshold add up over the whole experiment (ten in a thousand queries). This experiment-wide multiple-testing problem is not covered by the per-query E-value (Section 1.4).

1.2.7 Homology and the rest of this chapter

The remainder of this chapter follows from the ideas above. To measure similarity we need *alignments* and a *scoring scheme* grounded in how proteins evolve (Section 1.3). To use those scores to infer homology, we need to know their distribution for unrelated sequences, and to apply them to databases of hundreds of millions of sequences we need fast *heuristics* (Section 1.4). To push detection into the twilight zone we pool information from many homologues into *profiles* and *hidden Markov models* (Section 1.6). Because homology is often a property of domains rather than of whole proteins, we need *domain databases* (Section 1.9). The newest approach uses *protein language models*, trained on hundreds of millions of sequences. Their embeddings can, empirically, recover many family and structural relationships, although they are not themselves a test of common ancestry (Section 1.10).

See also: further explanation in Section D.1.1; related Wikipedia articles in Section E.1.1.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Homology (biology), Sequence homology, Gene, Protein and Convergent evolution (CC BY-SA 4.0, retrieved 2 October 2026).

1.3 Dynamic programming and sequence alignment

An alignment arranges two sequences in rows so that corresponding residues stand in the same column. Gaps (written −) are inserted where one sequence has gained or lost residues relative to the other. For homologous sequences, the intended interpretation is evolutionary: each column groups residues believed to derive from the same ancestral position, and the alignment is a *hypothesis* about the substitutions, insertions and deletions that separate the two sequences since their common ancestor (1). The algorithms themselves will align any two sequences, related or not; whether the alignment means anything biologically is a separate, statistical question. To choose between the astronomically many possible alignments we need two things:

1. a **scoring scheme**, which assigns a number to each alignment so that biologically plausible alignments score higher than implausible ones; and

2. an **algorithm** that finds the alignment with the best score without enumerating them all.

This section deals with both. The algorithm is *dynamic programming*, which finds the optimal alignment exactly, in time proportional to the product of the sequence lengths. The scoring scheme consists of a *substitution matrix*, which scores each pair of aligned residues, and a *gap penalty*, which scores insertions and deletions. Everything in this section is "exact": given the scores, the result is guaranteed to be optimal. The heuristics that trade some of this guarantee for speed are covered in Section 1.4.

1.3.1 A short history

The mathematical idea of dynamic programming was developed by **Richard Bellman** at the RAND Corporation in the early 1950s, for multistage decision problems in operations research (Bellman 1954). Bellman later explained that he chose the name partly because "programming" (in the sense of planning) and "dynamic" sounded impressive and were hard to object to politically. The same principle was then discovered independently in many fields that compare sequences:

- **1965–1966**: Vladimir Levenshtein defined the *edit distance* between two strings, the minimum number of insertions, deletions and substitutions needed to turn one into the other.
- **1968**: Taras Vintsyuk used dynamic programming to align speech signals (the forerunner of *dynamic time warping*).
- **1970**: **Saul Needleman and Christian Wunsch** published "A general method applicable to the search for similarities in the amino acid sequence of two proteins" (Needleman and Wunsch 1970). This was the first application of dynamic programming to biological sequences and the origin of *global* alignment. Their original formulation scored matches only and did not penalise gaps, and as published it required time proportional to the cube of the sequence length.
- **1972–1976**: David Sankoff, Peter Sellers, Michael Waterman, Temple Smith and William Beyer put the method on a rigorous footing: Sellers showed how to compute an evolutionary *distance* between sequences, and Waterman, Smith and Beyer (1976) introduced general gap penalties. The same algorithm for edit distance was published in computer science by Wagner and Fischer (1974).
- **1975**: Dan Hirschberg showed how to compute the longest common subsequence in linear space, a trick later adapted to alignment.
- **1978**: Margaret Dayhoff published the **PAM** substitution matrices (Dayhoff et al. 1978), the first scoring scheme derived from observed evolution.
- **1981**: **Temple Smith and Michael Waterman** published the *local* alignment algorithm (Smith and Waterman 1981), in a paper of only three pages. Local alignment finds the best-matching *segments* of two sequences and is the basis of all database searching.
- **1982**: Osamu Gotoh showed that alignment with *affine* gap penalties (a cost for opening a gap plus a smaller cost for extending it) can be done in quadratic time (Gotoh 1982).
- **1988**: Eugene Myers and Webb Miller combined Gotoh's algorithm with Hirschberg's linear-space idea (Myers and Miller 1988).
- **1990–1992**: Samuel Karlin and Stephen Altschul derived the statistics of local alignment scores (Karlin and Altschul 1990); this also explained what substitution matrices mean. Steven and Jorja Henikoff published the **BLOSUM** matrices (Henikoff and Henikoff 1992), still the default for protein searches.

Since then, the algorithms themselves have changed little. The main developments have been in speed — vectorised (SIMD) and GPU implementations of Smith–Waterman, and above all the heuristics of Section 1.4 — and in scoring, where position-specific profiles (Section 1.6) replace a single substitution matrix.

1.3.2 What is dynamic programming?

Dynamic programming solves a problem by breaking it into smaller, *overlapping* subproblems, solving each subproblem once, storing its solution in a table and building the solution of larger problems from the stored solutions. It applies when a problem has two properties:

- **Optimal substructure** (Bellman's *principle of optimality*): an optimal solution to the whole problem contains optimal solutions to its subproblems.
- **Overlapping subproblems**: the same subproblems recur many times, so storing their solutions (rather than recomputing them) saves time.

Sequence alignment has both properties. Consider an optimal alignment of the sequences $A = a_1 a_2 \dots a_n$ and $B = b_1 b_2 \dots b_m$. Its last column can only be one of three things: a_n aligned to b_m, a_n aligned to a gap, or b_m aligned to a gap. Whichever it is, the rest of the alignment must itself be an *optimal* alignment of the remaining prefixes — otherwise we could replace it by a better one and improve the whole. So the best score for the prefixes $a_1 \dots a_i$ and $b_1 \dots b_j$ can be computed from the best scores of three smaller pairs of prefixes, and the problem can be solved by filling in a table of $(n+1) \times (m+1)$ prefix scores.

Why is this necessary? The number of possible alignments grows exponentially. Two sequences of length n can be aligned in

$$\binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

ways (counting alignments that differ only in the order of adjacent gaps in the two sequences as the same) (Durbin et al. 1998). For two proteins of 300 residues this is about 10^{179} alignments, far more than could ever be enumerated. Dynamic programming finds the best one in about 9×10^4 steps.

A useful picture is the **alignment graph** (or *edit graph*): a grid with one sequence along the top and the other down the side. Every alignment is a path from the top-left to the bottom-right corner. A diagonal step aligns two residues, a horizontal step aligns a residue of the top sequence to a gap, and a vertical step aligns a residue of the side sequence to a gap. Finding the best alignment is the same as finding the highest-scoring path through this grid — a *longest-path problem in a directed acyclic graph*, which is the general form of almost every dynamic programming algorithm in bioinformatics (including the Viterbi algorithm for hidden Markov models, Section 1.6).

1.3.3 Global alignment: the Needleman–Wunsch algorithm

Global alignment aligns two sequences over their entire length. It is appropriate when the sequences are expected to be homologous along their whole length, for example two orthologous proteins of similar length with the same domain structure.

Let $s(a, b)$ be the score for aligning residue a with residue b (taken from a substitution matrix), and let each gap position cost d (a *linear* gap penalty; affine gaps are treated below). Define $F(i, j)$ as the score of the best alignment of the prefixes $a_1 \dots a_i$ and $b_1 \dots b_j$. Then

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(a_i, b_j) & \text{align } a_i \text{ with } b_j \\ F(i-1, j) - d & a_i \text{ aligned to a gap} \\ F(i, j-1) - d & b_j \text{ aligned to a gap} \end{cases}$$

with the boundary conditions $F(0,0) = 0$, $F(i,0) = -i\,d$ and $F(0,j) = -j\,d$ (a prefix aligned entirely to gaps). The table is filled row by row, so that the three cells needed for each new cell are always available. The optimal global score is $F(n, m)$.

```
NEEDLEMAN-WUNSCH(A[1..n], B[1..m], s, d)
    F[0,0] ← 0
    for i ← 1 to n:  F[i,0] ← −i·d
    for j ← 1 to m:  F[0,j] ← −j·d
    for i ← 1 to n:
        for j ← 1 to m:
            F[i,j] ← max( F[i−1,j−1] + s(A[i],B[j]),    # diagonal
                          F[i−1,j]   − d,               # up: gap in B
                          F[i,j−1]   − d )              # left: gap in A
    return F[n,m]
```

Traceback. The table gives the optimal *score*; the alignment itself is recovered by tracing back from $F(n, m)$ to $F(0, 0)$. At each cell we ask which of the three cases produced the value (either by storing a pointer while filling the table or by recomputing the three candidates), and step to that cell. A diagonal step emits the column $\binom{a_i}{b_j}$, a vertical step emits $\binom{a_i}{-}$ and a horizontal step $\binom{-}{b_j}$; the alignment is read off in reverse. When two or three cases give the same maximum, there are several **co-optimal alignments** with the same score. This is common, and it is a reminder that the optimal alignment is not necessarily the *correct* (biologically true) one.

Worked example: global alignment

Align GCATGCG and GATTACA with match +1, mismatch −1 and gap −1. Each cell holds $F(i, j)$; the blue cells and arrows show one traceback from the bottom-right corner (orange).

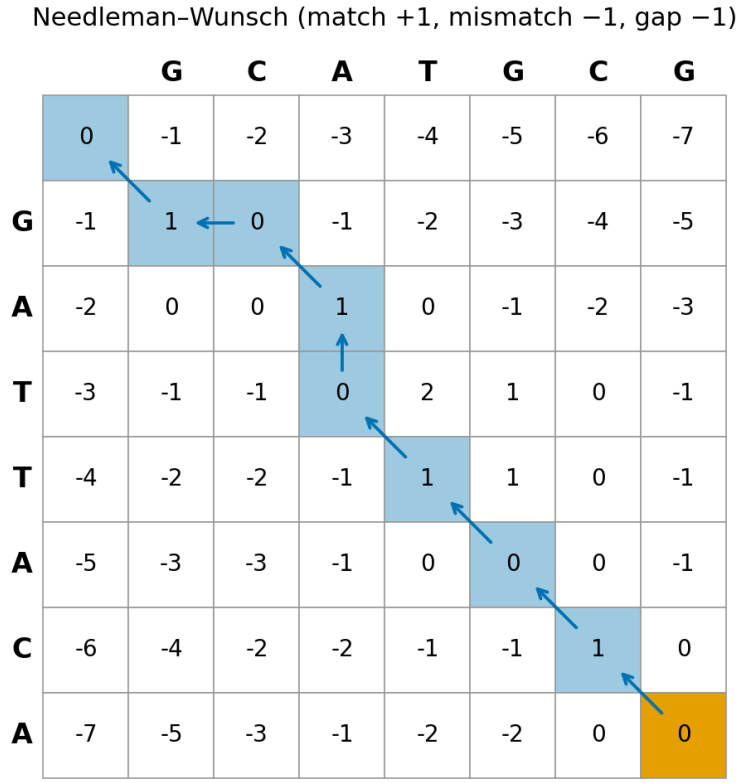


Figure 1.2: Needleman–Wunsch matrix for GCATGCG (top) and GATTACA (side), with one optimal traceback.

Needleman–Wunsch (global; match +1, mismatch −1, gap −1). Optimal score 0; there are 3 co-optimal alignments, one of which is:

```
GCA-TGCG
|  |  |  |
G-ATTACA
```

Check one cell: the cell for G (side) against G (top) is $\max(0+1,\ -1-1,\ -1-1) = 1$, coming from the diagonal.

1.3.4 Local alignment: the Smith–Waterman algorithm

Most real problems are *local*. Proteins are built from domains that can be shared between otherwise unrelated proteins (Section 1.9), a query may be a fragment, and distantly related sequences may be similar only in their conserved core. A global alignment would force the dissimilar parts to be aligned too, and their noise could swamp the signal of the similar part. **Local alignment** instead finds the pair of *segments* (one from each sequence) with the highest alignment score.

Smith and Waterman's insight was that only two changes to the global algorithm are needed (Smith and Waterman 1981):

1. **A fourth option of zero.** If the best alignment ending at (i, j) has a negative score, it is better to start a new alignment there:

$$H(i,j) = \max \begin{cases} 0 & \text{start a new local alignment} \\ H(i-1, j-1) + s(a_i, b_j) & \\ H(i-1, j) - d & \\ H(i, j-1) - d & \end{cases}$$

   with $H(i, 0) = H(0, j) = 0$.

2. **A different traceback.** The best local alignment can end anywhere, so the traceback starts from the *highest* value anywhere in the matrix and stops when it reaches a cell with value zero.

Because the matrix never goes below zero, a region of poor similarity "resets" the score, and the algorithm effectively finds the best-scoring diagonal stretch, gaps included. Further, non-overlapping local alignments can be found by declumping the matrix and repeating the traceback (Waterman and Eggert, 1987), which is how programs report several alignments between the same pair of sequences.

A condition on the scoring scheme. Local alignment only makes sense if the *expected* score of aligning two random residues is negative:

$$\sum_{a,b} p_a\, p_b\, s(a,b) < 0,$$

where p_a are the background frequencies of the residues, and at least one score is positive. If the expected score were positive, any long random alignment would accumulate a positive score, and the "local" alignment would simply extend over the full length of both sequences — the algorithm would behave like global alignment. All standard substitution matrices satisfy this condition, for a reason explained under *Substitution matrices* below.

Worked example: local alignment

Align TGTTACGG and GGTTGACTA with match +3, mismatch −3 and a linear gap penalty of 2 per position. Negative values never appear; the traceback starts at the maximum (13, orange) and stops at the first zero.

Smith–Waterman (match +3, mismatch −3, gap −2)

		G	G	T	T	G	A	C	T	A
	0	0	0	0	0	0	0	0	0	0
T	0	0	0	3	3	1	0	0	3	1
G	0	3	3	1	1	6	4	2	1	0
T	0	1	1	6	4	4	3	1	5	3
T	0	0	0	4	9	7	5	3	4	2
A	0	0	0	2	7	6	10	8	6	7
C	0	0	0	0	5	4	8	13	11	9
G	0	3	3	1	3	8	6	11	10	8
G	0	3	6	4	2	6	5	9	8	7

Figure 1.3: Smith–Waterman matrix for TGTTACGG (side) and GGTTGACTA (top). The best local alignment contains one gap.

Smith–Waterman (local; match +3, mismatch −3, gap −2). Highest cell 13; the traceback stops at the first zero and gives the local alignment:

```
GTT-AC
||| ||
GTTGAC
```

The alignment score is $5 \times 3 - 2 = 13$: five matches and one gap position.

1.3.5 Other alignment modes

Between fully global and fully local there are useful intermediate forms, all obtained by changing the boundary conditions and the cells where the traceback starts and stops:

- **Semi-global ("glocal" or end-gap-free) alignment** aligns all of one sequence to part of another — for example a short read to a genome, or a domain to a full-length protein. Leading and trailing gaps in the longer sequence are not penalised: the first row is initialised to zero, and the traceback starts from the best cell in the last row.

- **Overlap alignment** allows free end gaps in both sequences, so that the end of one sequence is aligned to the start of the other. It is used to detect overlaps between reads in genome assembly (Chapter 5).
- **Banded alignment** restricts the computation to a band of width k around a diagonal. This reduces the cost to $O(k\,n)$ and is exact if the optimal path is known to stay within the band, as when two sequences are already known to be very similar. Banded dynamic programming is used to refine the hits of heuristic methods (Section 1.4).

1.3.6 Edit distance: maximising similarity or minimising distance

Alignment can be framed in two equivalent ways. **Similarity** scores are maximised: matches score positively and mismatches and gaps negatively, as above. **Distances** are minimised: each operation has a non-negative cost. The simplest distance is the *Levenshtein* or *edit distance*, the minimum number of single-character substitutions, insertions and deletions needed to transform one string into the other. It is computed by the same recursion with min instead of max, a cost of 1 for each mismatch or gap and 0 for a match. For global alignment with suitable parameters, the two formulations give the same optimal alignments. Local alignment, however, has no natural distance formulation — a distance can always be made smaller by aligning shorter segments — which is one reason why bioinformatics works with similarity scores.

1.3.7 Gap penalties

Insertions and deletions (*indels*) arise from replication slippage, unequal crossing-over, transposition and other events, and one event often inserts or deletes several residues at once. A good gap penalty should therefore make one gap of length 3 cheaper than three separate gaps of length 1. Several forms are used; let $w(k)$ be the penalty for a gap of length k:

Gap penalty	$w(k)$	Comment
Linear	$k\,d$	Simplest; every gap position costs the same, so long gaps are over-penalised and gaps tend to be scattered.
Constant	g	Gap length does not matter; sometimes used for genomic DNA.
Affine	$g_o + (k-1)\,g_e$ with $g_o > g_e$	A cost for opening a gap plus a smaller cost for each extension. The standard choice.
Convex (e.g. logarithmic)	$g_o + g_e \log k$	Each additional gap position costs less than the previous one; fits observed indel length distributions better, but is more expensive to compute.
General	any function	Needs $O(n^2 m)$ time, as in the original Smith–Waterman paper.

(Some programs, such as BLAST, define the affine penalty as $g_o + k\,g_e$, counting the extension cost for every position including the first; the parameters are then not directly comparable between programs.)

Typical affine parameters for proteins are an opening penalty of about 10–12 and an extension penalty of 0.5–2 in units of the substitution matrix; BLAST with BLOSUM62 uses 11 and 1 by default.

The difference matters in practice. With a linear gap penalty the alignments of `TACGGGCCCGCTAC` and `TAGCCCTATCGGTCA` contain several isolated gap positions; with an affine penalty the same sequences align with a few longer gaps — fewer, biologically more plausible, indel events.

1.3.7.1 Affine gaps in quadratic time: Gotoh's algorithm

With a general gap function, each cell must consider gaps of every possible length ending there, which gives $O(n^2 m)$ time. Gotoh showed that for affine penalties it is enough to keep **three** matrices, one for each possible state of the last column (Gotoh 1982):

- $M(i, j)$: best score of an alignment of the prefixes that ends with a_i aligned to b_j;
- $I_x(i, j)$: best score of an alignment that ends with a_i aligned to a gap (a gap in B);
- $I_y(i, j)$: best score of an alignment that ends with b_j aligned to a gap (a gap in A).

With an affine penalty $w(k) = g_o + (k-1)g_e$ the recursions are

$$
\begin{aligned}
M(i,j) &= s(a_i, b_j) + \max\{M(i-1,j-1),\ I_x(i-1,j-1),\ I_y(i-1,j-1)\} \\
I_x(i,j) &= \max\{M(i-1,j) - g_o,\ \ I_x(i-1,j) - g_e\} \\
I_y(i,j) &= \max\{M(i,j-1) - g_o,\ \ I_y(i,j-1) - g_e\}
\end{aligned}
$$

and the score of the best alignment is the maximum of the three matrices in the final cell. (This version forbids a gap in A to be directly followed by a gap in B, which is never optimal when s is not too negative.) Opening a gap means leaving the match state; extending it means staying in the gap state. Each cell is still computed in constant time, so the cost remains $O(nm)$. The local version adds the zero option to M. Gotoh's original traceback procedure contained an error that could miss the optimal alignment in some cases; this was pointed out and corrected by Altschul and Erickson (1986).

The three matrices can be seen as the three states of a small finite-state machine — match, insertion, deletion — with transitions whose costs are the gap-opening and gap-extension penalties. This is exactly the structure of a *pair hidden Markov model*, in which the scores become log-probabilities (Durbin et al. 1998). Profile HMMs (Section 1.6) extend the same idea to whole families of sequences.

1.3.8 Time and space complexity

Filling one cell takes constant time, so all variants above need $O(nm)$ **time**. Storing the full matrix for the traceback needs $O(nm)$ **space**. For two proteins this is trivial, but for two sequences of 100,000 residues it already amounts to 10^{10} cells.

Linear space. The *score* alone needs only the previous row, so it can be computed in $O(\min(n, m))$ space. To recover the *alignment* in linear space, Hirschberg's divide-and-conquer method is used:

1. Compute the scores of the forward alignments of the top half of A against all prefixes of B, and the scores of the reverse alignments of the bottom half of A against all suffixes of B, keeping only the last row of each.
2. The optimal path must cross the middle row of A at the column j where the sum of the forward and reverse scores is maximal. This splits the problem into two independent subproblems — top-left and bottom-right of that point.
3. Recurse on each half.

The total work is about twice that of the ordinary algorithm, still $O(nm)$, while the space is $O(n+m)$. Myers and Miller extended this to affine gap penalties (Myers and Miller 1988), and it is the basis of space-efficient alignment tools.

Faster than quadratic? The "Four Russians" technique reduces the time to $O(nm/\log n)$ for small alphabets, but has little practical importance. More importantly, there is strong theoretical evidence that edit distance cannot be computed in strongly subquadratic time in the worst case. In practice, Smith–Waterman is accelerated by computing many cells at once: SIMD instructions on ordinary processors (as in the `SSEARCH` program of the FASTA package and in the SWIPE and Parasail libraries) and GPUs reach billions of cell updates per second.

Even so, exact local alignment of every query against a large database is expensive. UniProt contains roughly 10^{11} residues, so comparing one 300-residue protein against all of it requires about 3×10^{13} cell updates — minutes to hours on a modern computer, for a *single* query. Comparing all proteins of a newly sequenced genome, or a metagenome with millions of proteins, against the database would take years. This is the motivation for the heuristic methods of Section 1.4.

1.3.9 Substitution matrices

The algorithm guarantees the optimal alignment *for a given scoring scheme*; whether that alignment is biologically meaningful depends entirely on the scores. For DNA, simple schemes are often adequate. For proteins, the 20 amino acids differ greatly in size, charge, polarity and other properties, and some substitutions (for example isoleucine to valine) are accepted by evolution far more often than others (for example tryptophan to glycine). A **substitution matrix** is a 20×20 table of scores $s(a, b)$ that captures these differences.

1.3.9.1 Scores as log-odds ratios

Substitution scores are best understood as **log-odds scores** comparing two hypotheses about an aligned pair of residues a and b:

- the *match* (homology) model, in which a and b are derived from a common ancestral residue and occur together with **target frequency** q_{ab}, estimated from alignments of true homologues; and
- the *random* model, in which a and b are unrelated and occur together with probability $p_a p_b$, the product of their background frequencies.

The score is the logarithm of the ratio of these probabilities,

$$s(a,b) = \frac{1}{\lambda} \log \frac{q_{ab}}{p_a\, p_b},$$

where λ is a scale factor chosen so that the scores can be rounded to convenient integers. Matrices are usually given in *bits* (logarithm base 2) or fractions of bits; BLOSUM62, for example, is in half-bit units, so a score of $+2$ means the pair is $2^{2/2} = 2$ times more likely under the homology model than by chance. A positive score means the pair occurs more often in true alignments than expected by chance, a negative score that it occurs less often, and zero that there is no information.

Because scores are logarithms of probability ratios, and positions in an ungapped alignment are treated as independent, the **total score** of an ungapped alignment is the log-likelihood ratio of the whole alignment:

$$S = \sum_{k} s(x_k, y_k) = \frac{1}{\lambda} \log \frac{P(\text{alignment} \mid \text{homologous})}{P(\text{alignment} \mid \text{random})}.$$

Dynamic programming thus finds the alignment that is most likely under the homology model *relative* to the random model. The same interpretation underlies profile scores and hidden Markov models (Section 1.6).

Two further consequences follow:

- **The expected score is always negative.** The expected score of a random pair is

$$\sum_{a,b} p_a p_b \, s(a,b) = -\frac{1}{\lambda} \sum_{a,b} p_a p_b \log \frac{p_a p_b}{q_{ab}} \le 0,$$

  because the sum on the right is a relative entropy (Kullback–Leibler divergence), which is never negative. Any log-odds matrix therefore automatically satisfies the requirement of local alignment.
- **Every matrix is a log-odds matrix.** Karlin and Altschul showed the converse (Karlin and Altschul 1990): for any scoring matrix with a negative expected score and at least one positive entry, there is a unique $\lambda > 0$ with $\sum_{a,b} p_a p_b e^{\lambda s(a,b)} = 1$, and $q_{ab} = p_a p_b e^{\lambda s(a,b)}$ are the target frequencies for which the matrix is optimal. Choosing a matrix is therefore always, implicitly, choosing what kind of alignments one expects to find. The same λ appears in the statistics of alignment scores (Section 1.4).

The **relative entropy**

$$H = \sum_{a,b} q_{ab} \log_2 \frac{q_{ab}}{p_a p_b}$$

is the average score per aligned position in true alignments, in bits. It measures how much information each aligned pair carries for distinguishing homology from chance (Altschul 1991). Matrices for closely related sequences have high relative entropy (each position is informative, so a short alignment suffices); matrices for distant relationships have low relative entropy (a longer alignment is needed to reach significance).

1.3.9.2 Scoring DNA

For nucleotides, a simple match/mismatch scheme is usually used, for example $+1/-2$ or $+2/-3$ for highly similar sequences and $+1/-1$ for more divergent ones (the ratio sets the implied target identity). Since transitions (purine–purine or pyrimidine–pyrimidine changes, A↔G and C↔T) are about twice as common as transversions in most genomes, they can be given a less negative score. For protein-coding genes, alignment of the translated sequences is far more sensitive than DNA alignment, because the genetic code is degenerate and synonymous changes saturate quickly.

1.3.9.3 PAM matrices

The first substitution matrices derived from evolutionary data were Margaret Dayhoff's **PAM** matrices (Dayhoff et al. 1978). A *point accepted mutation* is the replacement of one amino acid by another that has been accepted by natural selection, that is, has become fixed in the population. Dayhoff's approach was explicitly evolutionary:

1. **Data.** 1572 accepted substitutions were counted on the branches of phylogenetic trees reconstructed for 71 families of closely related proteins (at least 85% identical), so that each observed difference could be assumed to be a single substitution.
2. **Mutation probability matrix.** From the counts, together with the *mutability* of each amino acid (how often it changes relative to how often it occurs) and the amino-acid frequencies, a matrix M_1 of transition probabilities was computed, scaled so that on average 1% of the amino acids change. This is **PAM1**: one accepted point mutation per 100 residues.

3. **Extrapolation.** Evolution was modelled as a Markov chain, so the matrix for a longer evolutionary distance is a power of M_1: $M_n = M_1^n$. PAM250 corresponds to 250 accepted mutations per 100 residues — many positions have changed several times — and to about 20% sequence identity.
4. **Log-odds scores.** The PAMn score is $s(i,j) = 10 \log_{10}\big(M_n(i,j)/f_i\big)$, the probability that j has become i after n PAM units relative to the background frequency of i.

Dayhoff found that asparagine, serine and aspartate are among the most mutable amino acids, and cysteine and tryptophan the least, in line with their special structural roles. The main weakness of PAM is the extrapolation: the substitution pattern of very similar sequences, which is dominated by single-nucleotide changes in the codon, is assumed to hold over long distances. The data set was also small by modern standards. Later matrices of the same type, such as JTT (1992) and WAG (2001), were estimated from much larger data sets and are used as substitution models in phylogenetics (Chapter 4).

1.3.9.4 BLOSUM matrices

Steven and Jorja Henikoff took a different, empirical approach (Henikoff and Henikoff 1992). Instead of extrapolating from close relatives, they counted substitutions directly in conserved regions of distantly related proteins:

1. **Data.** The BLOCKS database contains ungapped multiple alignments ("blocks") of the most conserved regions of protein families — about 2000 blocks from more than 500 families in the original work.
2. **Clustering.** To avoid over-counting closely related sequences, the sequences in each block that are more than r% identical are clustered, and each cluster contributes with a total weight of one sequence.
3. **Counting.** All pairs of amino acids in each column, between sequences in different clusters, are counted. This gives the observed pair frequencies q_{ab}, and from them the background frequencies p_a.
4. **Log-odds scores.** $s(a,b) = 2 \log_2\big(q_{ab}/(p_a p_b)\big)$, rounded to the nearest integer (half-bit units).

The number in the name is the clustering threshold: **BLOSUM62** is computed from blocks in which sequences more than 62% identical were clustered together. A higher number means less clustering, more similar sequences contribute and the matrix is suited to closely related sequences; a lower number suits distant relationships. The numbering therefore runs in the *opposite* direction to PAM. For PAM, a *higher* number means a greater evolutionary distance (more accepted mutations); for BLOSUM, a *lower* number is intended for more divergent sequences. BLOSUM80 corresponds roughly to a low PAM number, BLOSUM45 to a high one. The two numbers measure different things (an extrapolated distance and a clustering threshold) and cannot be compared directly.

In comparisons with PAM, BLOSUM matrices performed better in database searches, especially for distant relationships, and BLOSUM62 became the default of BLAST and most other protein search programs. Its relative entropy is about 0.70 bits per position, similar to PAM160; BLOSUM45 (about 0.38 bits) is comparable to PAM250. The derivation of BLOSUM62, and why it works so well, is explained in a short and very readable note by Eddy (2004b). A curious footnote is that the program originally used to compute BLOSUM62 contained errors, discovered only in 2008. The "correct" matrix turned out to perform slightly *worse* in searches than the miscalculated one, which is still the standard — a reminder that the best matrix is ultimately an empirical question.

1.3.9.5 Choosing a matrix

Because each matrix is optimal for a particular evolutionary distance — its implicit target frequencies — the best matrix depends on what one is looking for (Altschul 1991):

Purpose	Typical matrices
Very similar sequences, short queries or short peptides	PAM30, PAM70, BLOSUM80
General database searching	BLOSUM62 (the usual default)
Distant relationships, long alignments	BLOSUM45, PAM250

A "deep" matrix (low relative entropy) can detect distant homologues but needs long alignments to give significant scores, and it tends to extend alignments into unrelated regions. A "shallow" matrix (high relative entropy) gives short, accurate alignments of close relatives but misses distant ones. Some programs therefore search with several matrices, and modern programs adjust the matrix to the amino-acid composition of the sequences being compared (*composition-based statistics*), which matters for proteins with unusual compositions such as transmembrane or low-complexity proteins.

1.3.9.6 Limitations of a single substitution matrix

A substitution matrix assumes that the same substitution pattern applies at every position of every protein. In reality, a buried hydrophobic position, a catalytic residue and a surface loop evolve very differently. Position-specific scores, estimated from a multiple alignment of the family, capture this variation and are the basis of profiles and profile HMMs (Section 1.6); environment-specific matrices, which depend on the structural context of each residue, are used in fold recognition (Chapter 9).

1.3.10 From scores to significance

An optimal alignment always exists, even between two random sequences. The scores of optimal *local* alignments between unrelated sequences are not normally distributed but follow an **extreme-value (Gumbel) distribution**, because each score is itself the maximum over many possible alignments. Karlin and Altschul derived its parameters for ungapped alignments (Karlin and Altschul 1990): the expected number of alignments with score at least S between random sequences of lengths m and n is

$$E = Kmne^{-\lambda S},$$

with λ the same scale factor as above and K a second constant that depends on the scoring system. For gapped alignments the parameters are estimated by simulation. These *E-values*, and their use in database searches, are treated in Section 1.4. Global alignment scores have no comparably simple theory, and their significance is usually assessed by shuffling the sequences.

1.3.11 Summary

- Dynamic programming finds the optimal alignment of two sequences under a given scoring scheme exactly, in $O(nm)$ time, by building optimal alignments of prefixes from optimal alignments of shorter prefixes.
- Global (Needleman–Wunsch), local (Smith–Waterman), semi-global and overlap alignment differ only in their boundary conditions, the zero option and where the traceback starts and stops.
- Affine gap penalties are handled in quadratic time with three matrices (Gotoh), and alignments can be recovered in linear space (Hirschberg, Myers–Miller).
- Substitution scores are log-odds ratios of target and background frequencies. Every matrix implies a set of target frequencies and is best suited to one evolutionary distance; PAM extrapolates from close relatives, BLOSUM counts directly in conserved blocks.

- The optimal alignment is only as good as the scoring scheme, and an optimal alignment is not necessarily the true one. Whether a score indicates homology is a statistical question (Section 1.4), and for sequence identities below about 25–30% the answer is often uncertain (1).

See also: further explanation in Section D.1.2; related Wikipedia articles in Section E.1.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Dynamic programming, Sequence alignment, Needleman–Wunsch algorithm, Smith–Waterman algorithm, Hirschberg's algorithm, Gap penalty, Substitution matrix, BLOSUM and Point accepted mutation (CC BY-SA 4.0, retrieved 2 October 2026). The worked examples follow those in the Needleman–Wunsch and Smith–Waterman articles.

1.4 Heuristic search methods

Dynamic programming (Section 1.3) is guaranteed to find the best alignment for a given scoring scheme, but it is expensive. Comparing a query of length m with a database of total length N takes time proportional to mN. For a typical protein of 300 residues searched against UniProtKB, with a few hundred million sequences, that means tens of trillions of dynamic-programming cells for every query. Even with modern vectorised implementations, which can fill billions of cells per second, this takes hours. Databases have also grown much faster than computers have become faster: sequence databases have roughly doubled every two years since the 1980s, and metagenomic projects now produce billions of protein sequences.

Practical database searching therefore uses **heuristics**: methods that first find short, promising regions of similarity quickly, and then spend the expensive dynamic-programming effort only on those regions. The price is that a heuristic can miss alignments that the full Smith–Waterman algorithm would have found. The heuristic lies mainly in the *filter*, the decision which sequences and regions are examined at all. The alignments that are finally reported are computed by dynamic programming, usually restricted to a band or by X-drop extension around the seed. For clear homologues they are therefore usually identical, or very close, to the Smith–Waterman alignment, although an X-drop extension can stop early and report a shorter alignment. The art of designing a search tool lies in choosing a filter that discards almost all unrelated sequences while keeping almost all true homologues. Equally important is the second half of the problem: deciding which of the reported hits are significant. This section covers both, from FASTA and BLAST to the tools used today for searching billions of sequences and structures, and the exact-matching indexes used for mapping sequencing reads.

1.4.1 FASTA

The first widely used heuristic was developed by David Lipman and William Pearson. FASTP, for protein searches, was published in 1985 (Lipman and Pearson 1985). Its successor FASTA, which extended the approach to nucleotide and translated protein–DNA searches, followed in 1988 (Pearson and Lipman 1988). The name lives on in the FASTA file format, which was introduced with the program and is now the standard way to store sequences.

FASTA builds on a simple observation. If two sequences are related, their alignment usually contains several short stretches of identical residues, and these lie on the same *diagonal* of the dot plot (the same offset between the two sequences). FASTA proceeds in four steps:

1. **Find word hits.** All identical words of length `ktup` (typically 1 or 2 for proteins, 4–6 for DNA) shared by the query and the database sequence are found with a lookup table. The time needed is proportional to the sequence lengths, not to their product.
2. **Find the best diagonal regions.** Hits on the same diagonal are grouped, and each diagonal region is scored by the number of hits minus a penalty for the distance between them. The ten best regions are kept and rescored with a substitution matrix (originally PAM250, now BLOSUM50 by default), which allows conservative substitutions and runs of identities shorter than `ktup` to contribute. The best single region gives the score `init1`.
3. **Join regions.** Regions on nearby diagonals are combined into one longer alignment containing gaps, at a cost of a joining penalty. The score of the best combination is `initn`.
4. **Optimise.** For the best-scoring database sequences, a banded Smith–Waterman alignment is computed in a narrow band (for example 32 residues wide) around the best diagonal. This gives the final score `opt`.

A smaller `ktup` makes the search more sensitive but slower. A distinctive feature of FASTA is how it estimates significance. Because a database search compares the query with many sequences, most of which are unrelated, the distribution of their scores can itself serve as the null model. FASTA fits an extreme-value distribution to the scores of the unrelated sequences and reports an expectation value for each hit (Pearson 2013). The FASTA package also contains SSEARCH, a full Smith–Waterman implementation that is often used as a gold standard for sensitivity.

1.4.2 BLAST

The **Basic Local Alignment Search Tool** was published in 1990 by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers and David Lipman at the NCBI (Altschul et al. 1990). It became one of the most cited papers in all of science. BLAST combined two ideas. The first was a faster and more selective seeding strategy than FASTA's. The second, at least as important, was a rigorous statistical theory for local alignment scores, developed in parallel by Samuel Karlin and Altschul (Karlin and Altschul 1990) and described below.

1.4.2.1 The original algorithm (1990)

BLAST looks for **high-scoring segment pairs** (HSPs): pairs of equal-length segments, one from each sequence, aligned without gaps, whose score cannot be improved by extending or shortening them. The highest-scoring of these is the **maximal segment pair** (MSP). The original BLAST approximated the MSP score in three steps.

1. **Compile a word list.** The query is split into all overlapping words of length w (usually $w = 3$ for proteins). For each query word, BLAST lists all words of the same length that score at least T against it with the substitution matrix (BLOSUM62 by default). These are the *neighbourhood words*. For example, with BLOSUM62 the query word PQG scores 15 against PEG ($7 + 2 + 6$) and 12 against PQA ($7 + 5 + 0$). With $T = 13$, PEG is in the neighbourhood of PQG but PQA is not. The neighbourhood is what distinguishes BLAST from FASTA: seeds do not have to be identical, only similar, so they are much more informative. For DNA, BLAST instead uses exact matches of a longer word (by default $w = 11$).
2. **Scan the database.** The neighbourhood words are compiled into a finite-state automaton (or a lookup table), so that each database sequence can be scanned in a single pass for exact occurrences of any of them. Each occurrence is a *hit*.

3. **Extend the hits.** Each hit is extended in both directions without gaps, adding the substitution scores residue by residue. The extension stops when the running score has fallen more than X below the best score seen so far (the *X-drop* rule). The best segment found is reported as an HSP if its score is high enough.

The parameter T controls the trade-off between speed and sensitivity. A high T gives few neighbourhood words and few hits to extend, so the search is fast but may miss alignments that contain no high-scoring word. A low T gives the opposite.

1.4.2.2 Gapped BLAST and the two-hit method (1997)

The 1990 BLAST reported only ungapped alignments. A real alignment interrupted by indels therefore appeared as several separate HSPs, and weak alignments could be missed altogether. In 1997, Altschul and co-workers published gapped BLAST together with PSI-BLAST (Altschul et al. 1997), with two main improvements.

- **The two-hit method.** Most time in the original BLAST was spent extending hits that led nowhere. The authors observed that a biologically meaningful HSP is usually much longer than a single word, so it will typically contain *two* word hits on the same diagonal. In gapped BLAST, an ungapped extension is triggered only when two non-overlapping hits occur on the same diagonal within a distance A (by default 40 residues) of each other. This allowed T to be lowered, increasing sensitivity, while the number of extensions fell sharply.
- **Gapped extension.** If an ungapped HSP scores above a moderate threshold, a gapped alignment is computed by dynamic programming, starting from a central residue pair of the HSP and extending in both directions. The extension uses the X-drop rule: cells whose score has fallen more than X_g below the best score are not explored. This confines the computation to a small region around the seed rather than the whole matrix.

The result was a program about three times faster than the original, which reports gapped alignments. PSI-BLAST, the iterative profile search introduced in the same paper, is described in Section 1.6.

1.4.2.3 Low-complexity filtering

Regions of biased composition, such as stretches rich in a single amino acid, short tandem repeats or coiled coils, align with high scores to many unrelated sequences. Because they violate the assumptions of the statistical model, they produce many false positives with apparently excellent E-values. Such hits can be very strong and still biologically meaningless: two proteins that both contain a polyglutamine or a proline-rich stretch are not thereby homologous. Two remedies are used. *Masking* identifies low-complexity regions in the query, with SEG (Wootton and Federhen 1993) for proteins and DUST for DNA, and replaces them by X (protein) or N (DNA) so that they cannot seed alignments (*soft masking* excludes them from seeding only, but allows them in the extension). *Composition-based statistics* (Schäffer et al. 2001) instead adjust the scores of each alignment to the amino-acid composition of the two sequences (see below). Which of these is on by default differs between programs and versions (in current BLAST+, DUST masking is the default for blastn, and composition-based statistics rather than SEG for blastp), so the settings should be checked and reported. Masking is also a common reason why a search misses a hit that seems obvious.

1.4.2.4 The BLAST family of programs

The same engine is used in a family of programs that differ in what is compared with what:

Program	Query	Database	Typical use
blastp	protein	protein	find homologues of a protein
blastn	nucleotide	nucleotide	DNA/RNA similarity
megablast	nucleotide	nucleotide	very similar sequences (long exact seeds, $w = 28$); fast
blastx	nucleotide, translated in six frames	protein	find what protein a DNA fragment or transcript encodes
tblastn	protein	nucleotide, translated in six frames	find (unannotated) genes for a protein in genomes
tblastx	translated nucleotide	translated nucleotide	compare two DNA sequences at the protein level; slow

Translated searches are important because protein sequences diverge much more slowly than the DNA that encodes them. Two genes whose DNA is no longer recognisably similar, because of synonymous changes and codon usage, may still encode clearly similar proteins. In general, searching at the protein level is far more sensitive than at the DNA level. Translated searches are also how nucleotide data (genomes, transcripts, metagenomic reads) are connected to the protein databases, where most functional annotation lives. Two practical points follow from the translation. The search space is measured in amino-acid residues, so a translated nucleotide sequence contributes about a third of its length per reading frame. And a sequencing error that causes a frameshift splits a translated alignment into pieces in different frames; some tools, such as FASTX and FASTY in the FASTA package and DIAMOND's frameshift-aware mode, can align across such frameshifts.

1.4.2.5 BLAT and other genome-scale tools

BLAST builds its word index from the *query* and scans the database. For aligning millions of mRNAs or ESTs to a genome, the reverse is more efficient. BLAT, the BLAST-like alignment tool written by Jim Kent for the human genome project (Kent 2002), builds an index of non-overlapping k-mers of the whole genome, keeps it in memory, and then looks up the k-mers of each query. It is designed for finding very similar sequences (for example the same gene in the same or a closely related species) and is hundreds of times faster than BLAST for that task, but much less sensitive for distant homologues. The same idea, an index of the *reference*, underlies the read mappers discussed at the end of this section. Aligning whole genomes to each other, where rearrangements, duplications and repeats must be handled, needs further steps, described in Section 1.8.

1.4.3 The statistics of alignment scores

Every search reports a best hit, even for a random query. The crucial question is therefore not "how similar is the hit?" but "how often would we see a score this high by chance?" Answering it requires knowing the distribution of scores between unrelated sequences.

1.4.3.1 Why not a normal distribution?

The score of an alignment of two *fixed* segments is a sum of many substitution scores, and so by the central limit theorem it is approximately normally distributed. But a local alignment score is the **maximum** over all possible pairs of segments. The maximum of many random variables does not follow a normal distribution. Its distribution belongs to the family of **extreme value distributions**, and for local alignment scores it is of the **Gumbel** type. The Gumbel distribution has a much heavier right tail than a normal distribution: it decays exponentially, as $e^{-\lambda x}$, whereas the normal tail falls off as $e^{-x^2/2}$. Using a normal model, or a z-score, therefore greatly underestimates how often high scores occur by chance.

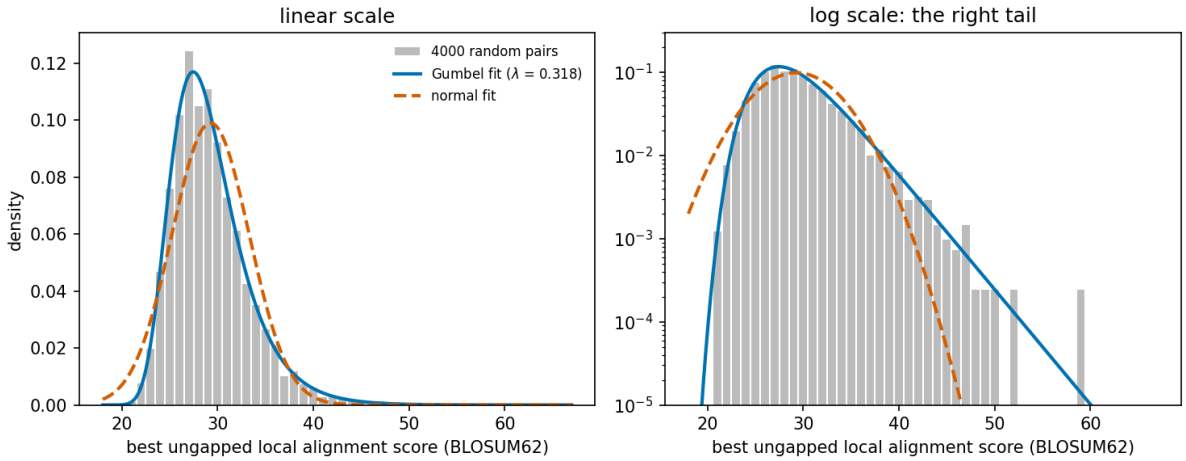


Figure 1.4: Distribution of the best ungapped local alignment score between pairs of random protein sequences (length 250, BLOSUM62). A Gumbel distribution (blue) fits the data, including the right tail; a normal distribution fitted to the same data (orange, dashed) underestimates the probability of high scores by orders of magnitude.

In this simulation (4000 pairs of random sequences of length 250, BLOSUM62, background frequencies of Robinson & Robinson), the fitted Gumbel distribution has $\lambda \approx 0.318$ and $K \approx 0.099$, close to the value $\lambda = 0.318$ predicted by solving $\sum_{ij} p_i p_j e^{\lambda s_{ij}} = 1$. For a score of 48, the Gumbel model gives $P(S \geq x) = 1.5 \times 10^{-3}$, whereas a normal distribution fitted to the same data gives 1.7×10^{-6}: the normal model underestimates how often high scores occur by chance.

1.4.3.2 Karlin–Altschul statistics

For *ungapped* local alignments, Karlin and Altschul derived the score distribution analytically (Karlin and Altschul 1990). The theory assumes:

- the residues of the two sequences are independent and drawn from background frequencies p_i;
- the expected score of aligning two random residues is negative, $\sum_{ij} p_i p_j s_{ij} < 0$ (otherwise the best "local" alignment would grow with the length of the sequences and become a global one);
- at least one score s_{ij} is positive.

Under these conditions, for two random sequences of lengths m and n that are long enough, the number of distinct HSPs with score at least S is approximately Poisson-distributed with mean

$$E = Kmne^{-\lambda S}.$$

This is the **E-value**: the number of alignments with score $\geq S$ expected by chance. Two parameters depend on the scoring system and the background frequencies. λ is the unique positive solution of

$$\sum_{i,j} p_i \, p_j \, e^{\lambda s_{ij}} = 1,$$

and acts as a scale that converts raw scores into natural units. K is a smaller correction factor for the fact that high-scoring alignments are not independent: an HSP is usually surrounded by overlapping, slightly lower-scoring ones. For BLOSUM62 with standard amino-acid frequencies, $\lambda \approx 0.318$ and $K \approx 0.13$ for ungapped alignments. Equivalently, the maximal score follows a Gumbel distribution,

$$P(S \geq x) \approx 1 - \exp\!\left(-e^{-\lambda(x-\mu)}\right), \qquad \mu = \frac{\ln(Kmn)}{\lambda}.$$

The theory also connects to substitution matrices (Section 1.3). Every scoring matrix with negative expected score is implicitly a log-odds matrix, $s_{ij} = \frac{1}{\lambda} \ln \frac{q_{ij}}{p_i p_j}$, for some *target frequencies* q_{ij}. These are the frequencies with which residue pairs occur in the alignments the matrix is best at finding.

1.4.3.3 E-values, P-values and bit scores

Raw scores depend on the scoring system, so they cannot be compared between searches that use different matrices. Normalising with λ and K gives the **bit score**

$$S' = \frac{\lambda S - \ln K}{\ln 2},$$

in terms of which the E-value takes a particularly simple form:

$$E = m n \, 2^{-S'}.$$

Every extra bit of score halves the E-value. The "search space" mn appears explicitly: in a database search, n is the total length of the database. Since the number of chance HSPs is Poisson-distributed, the probability of finding *at least one* with score $\geq S$ is the **P-value**

$$P = 1 - e^{-E}.$$

For small values, $P \approx E$. For large values they differ: $E = 3$ means that about three such hits are expected by chance, but P can never exceed 1. BLAST therefore reports E-values, which are easier to interpret over the whole range.

1.4.3.4 Corrections and gapped alignments

Several refinements are needed in practice.

- **Edge effects.** An alignment cannot start near the end of a sequence and still be long. BLAST therefore uses *effective lengths*: m and n are shortened by the expected length of an HSP, $\ell \approx \ln(Kmn)/H$, where H is the relative entropy of the scoring system. For a database of many sequences, the correction is applied to each sequence.
- **Gapped alignments.** No exact theory exists for gapped alignments. Simulations show that their scores still follow a Gumbel distribution to a good approximation, but λ and K must be estimated empirically for each combination of matrix and gap penalties. For example, for BLOSUM62 with gap open 11 and extend 1, $\lambda \approx 0.267$ and $K \approx 0.041$. These precomputed values are built into BLAST, which is why only certain combinations of matrix and gap costs are allowed. Very low gap penalties make the score behave like a global one, and the statistics break down.
- **Composition.** Proteins with unusual amino-acid composition, such as transmembrane or low-complexity proteins, violate the assumption of standard background frequencies. Modern BLAST uses *composition-based statistics*, which rescale the scores, or adjusts the substitution matrix itself to the composition of the two sequences being compared (Schäffer et al. 2001).
- **Multiple HSPs.** When two sequences share several HSPs, for example around an indel, *sum statistics* combine them into a single, more significant E-value (Karlin and Altschul 1993).

A different approach to significance is to estimate the null distribution directly. FASTA fits the distribution of scores from the database search itself (Pearson 1998). HMMER uses probabilistic models whose score distributions can be derived or calibrated (Eddy 2008). Another approach is to re-align the hit after shuffling one of the sequences many times.

1.4.3.5 Interpreting E-values

- **The E-value depends on the database.** Because $E = mn\,2^{-S'}$ and n is the length of the whole database, the same alignment, with exactly the same raw score and bit score, gets a ten times larger E-value when it is found in a database ten times larger. The score measures the similarity of the two sequences; the E-value measures how surprising that score is in this particular search. This is not a flaw but the correct accounting for the number of comparisons: the E-value is a built-in correction for multiple testing *within one query's search*. For the same alignment score, searching a smaller, targeted database therefore gives a smaller E-value, so a borderline relationship can pass a fixed significance threshold that it would not pass in a much larger database (the search algorithm finds the same alignment either way). E-values from searches of different database releases are not directly comparable (bit scores are). As databases grow, a borderline hit can become non-significant without any change in its alignment.
- **Many queries are a separate problem.** The E-value does not account for the number of *queries*. When 10,000 proteins of a genome are each searched with a threshold of $E = 0.01$, about 100 chance hits are expected over the whole experiment. For genome-scale analyses the threshold must be chosen with this experiment-wide error in mind (see Section 7.4 for family-wise error and false discovery rates).
- **Thresholds.** For a single protein search, $E < 10^{-3}$ is usually safe evidence of homology, and many studies use $E < 10^{-5}$ or lower to be conservative. E-values around 0.01–1 are a "twilight" region that needs other evidence, for example reciprocal searches, profiles or structure.
- **Significance means homology, not identical function.** A significant E-value means that the similarity is very unlikely to arise by chance. The most plausible explanation is common ancestry (1). Homology is a yes/no statement, so a better E-value does not mean "more homologous", and it does not guarantee that the hit has the same function (Chapter 10).

- **Non-significance does not mean non-homology.** Many homologues have diverged too far to be detected by pairwise comparison. Profiles, profile HMMs (Section 1.6) and structure comparison detect many of them.
- **The model can fail.** Low-complexity regions, repeats, coiled coils and transmembrane segments can give spurious significant E-values. So can searches against databases full of near-identical sequences, where one real hit appears thousands of times.

1.4.4 Searching at the scale of billions of sequences

BLAST was designed when databases contained tens of thousands of sequences. Metagenomics now produces billions of protein sequences, and analyses such as annotating every read of a microbiome sample or building multiple alignments for structure prediction need millions of searches. A new generation of tools reaches BLAST-like sensitivity at speeds hundreds to thousands of times higher. They combine better filters with cache-efficient and vectorised (SIMD) implementations on modern hardware.

1.4.4.1 DIAMOND

DIAMOND was written by Benjamin Buchfink and Daniel Huson, first for aligning short metagenomic DNA reads against protein databases (a blastx task) (Buchfink et al. 2014), and later extended to sensitive protein searches at the scale of the whole tree of life (Buchfink et al. 2021). Its main ideas are:

- **A reduced alphabet.** Amino acids are grouped into 11 classes of similar residues, so that seeds can match across conservative substitutions.
- **Spaced seeds.** A seed is defined by a pattern in which some positions must match and others are ignored, for example `111011`. Spaced seeds are more sensitive than contiguous words of the same weight, because hits at different positions of a real alignment are less correlated.
- **Double indexing.** Instead of scanning the database for the words of each query, DIAMOND builds seed tables for *both* the query set and the database, sorts them, and walks through the two lists together. Matches are found in a single linear pass with very good use of the CPU cache. This makes DIAMOND particularly efficient when many queries are searched at once.

Seed hits are then extended by ungapped and gapped alignment much as in BLAST, and E-values are computed with BLAST-style statistics. In its most sensitive modes, DIAMOND reaches the sensitivity of blastp while being roughly 100 times faster or more (Buchfink et al. 2021).

1.4.4.2 MMseqs2

MMseqs2 (Many-against-Many sequence searching), developed by Martin Steinegger and Johannes Söding (Steinegger and Söding 2017), is now one of the most widely used tools for large-scale protein search and clustering. A search proceeds in three stages, each more expensive and more selective than the last.

1. ***k*-mer prefilter.** For each query, MMseqs2 generates all *similar* *k*-mers (k = 6 or 7): all words that score above a threshold against a *k*-mer of the query with the substitution matrix, exactly like BLAST's neighbourhood words but longer. These are looked up in an index of the target database. The key filter is to require **two consecutive *k*-mer matches on the same diagonal**, a generalisation of BLAST's two-hit method. The sensitivity parameter (`-s`) controls how many similar *k*-mers are generated: higher sensitivity means more *k*-mers, more candidate hits and a slower search.

2. **Ungapped alignment.** Candidates that pass the prefilter are scored by a fast vectorised ungapped alignment along the matching diagonal.
3. **Gapped alignment.** The remaining candidates are aligned with a vectorised Smith–Waterman algorithm, and E-values are computed.

Because the stages are separated, MMseqs2 can also run **iterative profile searches**, analogous to PSI-BLAST (Section 1.6): the hits of one iteration are turned into a position-specific scoring profile that is used as the query in the next. In this mode it is more sensitive than PSI-BLAST at a small fraction of the cost (Steinegger and Söding 2017). MMseqs2 also includes tools for taxonomic assignment of metagenomic sequences and for clustering.

1.4.4.3 Clustering huge sequence sets

Many applications need non-redundant databases, in which similar sequences are represented by a single *representative*. Examples are UniRef, which clusters UniProt at 100%, 90% and 50% sequence identity (Suzek et al. 2014), and the large metagenomic databases used for structure prediction. Clustering by all-against-all comparison scales as $O(N^2)$ in the number of sequences, which is impossible for billions of them. **Linclust** (Steinegger and Söding 2018) clusters in time proportional to N:

1. for each sequence, select a small number m of its k-mers, chosen by a hash function so that similar sequences tend to select the same k-mers;
2. group sequences that share a selected k-mer, and pick the longest sequence in each group as its centre;
3. align each sequence only with the centre of its group, and accept it as a member if the alignment meets the identity and coverage criteria.

Each sequence is compared with only a few centres, so the cost grows linearly. Linclust was used to cluster more than a billion metagenomic protein fragments. These clustered databases, searched with MMseqs2, are now used to build the multiple sequence alignments for protein structure prediction. ColabFold, for example, uses an MMseqs2 server to replace the slower HMMER/HHblits searches of the original AlphaFold2 pipeline, making structure prediction tens of times faster (Mirdita et al. 2022) (see Chapter 9).

1.4.5 Searching with structure: Foldseek

Protein structure is conserved much longer than sequence: proteins with less than 20% sequence identity, whose relationship is invisible to sequence search, often still have the same fold. Structure comparison can therefore detect homologues that sequence methods miss. But traditional structural aligners are slow, typically fractions of a second to seconds per pair. DALI aligns two structures by comparing their matrices of intramolecular Cα–Cα distances, so that residues with similar distance patterns to their neighbours are matched (Holm and Sander 1993). TM-align searches for the superposition and residue alignment that maximise the TM-score, a length-normalised measure of structural similarity (Zhang and Skolnick 2005). CE builds the alignment from fragment pairs with similar local geometry. That was acceptable for the PDB, with about 200,000 structures, but not for the more than 200 million predicted structures in the AlphaFold database (Chapter 2).

Foldseek (van Kempen et al. 2023), from the Steinegger and Söding groups, solves this by turning a structure into a *sequence* over a new alphabet, so that the fast sequence-search machinery of MMseqs2 can be reused.

- **The 3Di alphabet.** Each residue is described by the geometry of its interaction with its spatially closest residue: distances and angles between the two backbones and their neighbours. These descriptors are mapped to one of 20 discrete states by a *vector-quantised variational autoencoder* (VQ-VAE), a neural network trained so that the states are maximally conserved between structurally aligned residues of homologous proteins. A protein structure thus becomes a string of 3Di letters, one per residue. Unlike classical structural alphabets, which describe the local *backbone* conformation (such as helix or strand), 3Di describes *tertiary* interactions. These are more informative and better conserved.
- **A 3Di substitution matrix.** As for amino acids, a log-odds substitution matrix for 3Di states was derived from alignments of homologous structures. Foldseek scores alignments with the sum of the 3Di score and an amino-acid score, which improves sensitivity further.
- **MMseqs2 search.** Prefiltering, ungapped and gapped alignment then work exactly as for sequences. Alignments can optionally be refined with TM-align to give global structural superpositions and TM-scores.

Foldseek reduced search times by four to five orders of magnitude, while reaching most of the sensitivity of DALI and TM-align on benchmarks (van Kempen et al. 2023). Searching the entire AlphaFold database with a single structure now takes seconds to minutes. Combined with the clustering methods above, it was used to cluster the 214 million structures of the AlphaFold database into about 2.3 million non-singleton structural clusters. Many of these clusters have no known function, and some represent folds that had never been seen before (Barrio-Hernandez et al. 2023). Structure comparison is discussed further in Chapter 9.

1.4.6 Exact matching: suffix arrays and the Burrows–Wheeler transform

A different problem arises with high-throughput sequencing: mapping hundreds of millions of short reads, each 100–300 bases long and nearly identical to the reference, to a genome of three billion bases. Here the challenge is not finding distant homologues but finding near-exact matches extremely fast. The solution is to *index the reference* once, in a data structure that allows any pattern to be located in time that depends on the length of the pattern rather than of the genome.

1.4.6.1 Suffix trees and suffix arrays

A **suffix tree** of a text is a compressed tree containing all suffixes of the text. Every substring is a path from the root, so a pattern of length p can be found in $O(p)$ time. Suffix trees can be built in linear time, but they need 10–20 bytes per character, far too much for a mammalian genome. A **suffix array**, introduced by Manber and Myers in 1990 (Manber and Myers 1993), stores only the start positions of all suffixes, sorted lexicographically. A pattern is found by binary search in $O(p \log n)$ time. This needs less memory (4–8 bytes per character), but still about 12 GB for the human genome.

1.4.6.2 The Burrows–Wheeler transform

The **Burrows–Wheeler transform** (BWT) was invented by David Wheeler in 1983 and published with Michael Burrows in 1994 as the basis of a text compression algorithm, which is used in `bzip2`. The transform is defined as follows:

1. append a unique end-of-text symbol \$, which sorts before all other characters;
2. write down all rotations of the text;
3. sort the rotations lexicographically;
4. the BWT is the last column L of the sorted matrix.

Because the rows are sorted, the first column F is simply the sorted characters of the text. The rows of the matrix correspond one-to-one to the suffixes of the text, so the order of the rows is given by the suffix array. The BWT has two remarkable properties. First, it is reversible: the original text can be reconstructed from L alone. Second, it tends to group equal characters into runs, because characters followed by the same context end up next to each other. This makes it very compressible.

The key to reversing it, and to searching with it, is the **last-to-first (LF) mapping**. The i-th occurrence of a character c in L and the i-th occurrence of c in F are the *same* character of the text. This is because both sets of occurrences are ordered by the text that follows them. In addition, $L[i]$ is the character that precedes $F[i]$ in the text. Using these two facts, one can walk backwards through the text one character at a time.

1.4.6.3 The FM-index and backward search

Paolo Ferragina and Giovanni Manzini showed in 2000 that the BWT, together with two small auxiliary tables, forms a compressed full-text index: the **FM-index** (Ferragina and Manzini 2000). The tables are:

- $C[c]$: the number of characters in the text that are smaller than c. The rows starting with c form the interval $[C[c], C[c] + n_c)$, where n_c is the number of occurrences of c;
- $\mathrm{Occ}(c, k)$: the number of occurrences of c in the first k positions of L. It is stored at regular checkpoints, so it can be computed in constant time.

A pattern is located by **backward search**. We start with the interval of rows beginning with the last character of the pattern, and extend the match one character to the left at a time, using the LF mapping to compute the new interval. After p steps, the size of the interval is the number of occurrences of the pattern, and the suffix array (sampled to save memory) gives their positions. Counting takes $O(p)$ time, independent of the size of the genome, and the whole index for the human genome fits in a few gigabytes. A small example:

Step 1 — sort the rotations of `ACAACG$`. The rows of the sorted matrix are the suffixes of the text in lexicographic order, so the column *SA* (the start position of each suffix) is the suffix array. F is the first and L the last column; L is the BWT.

Row	Sorted rotation	F	L	SA
0	`$ACAACG`	`$`	`G`	6
1	`AACG$AC`	`A`	`C`	2
2	`ACAACG$`	`A`	`$`	0
3	`ACG$ACA`	`A`	`A`	3
4	`CAACG$A`	`C`	`A`	1
5	`CG$ACAA`	`C`	`A`	4
6	`G$ACAAC`	`G`	`C`	5

The BWT of `ACAACG$` is **`GC$AAAC`**. Note how equal characters tend to cluster in L; in a real genome this makes the BWT highly compressible.

Step 2 — the tables used by the FM-index. $C[c]$ is the number of characters in the text that are smaller than c; $\mathrm{Occ}(c, k)$ is the number of times c occurs in the first k positions of L.

c	`$`	`A`	`C`	`G`
$C[c]$	0	1	4	6

Step 3 — backward search for `AC`. The rows whose rotation starts with a given string always form one contiguous interval $[s, e)$. We start with the interval for the last character of the pattern and extend it one character to the left at a time with $s \leftarrow C[c] + \mathrm{Occ}(c, s)$ and $e \leftarrow C[c] + \mathrm{Occ}(c, e)$.

Step	Character	Suffix matched	Interval $[s, e)$	Rows
1	`C`	`C`	$[4, 6)$	4, 5
2	`A`	`AC`	$[2, 4)$	2, 3

The final interval contains 2 rows, so `AC` occurs 2 times. Looking up these rows in the suffix array gives the positions 0, 3 (0-based) in `ACAACG$`. Each step costs a constant number of table look-ups, so counting takes $O(p)$ time for a pattern of length p, independent of the length of the text.

1.4.6.4 Use in read mapping

The FM-index is the basis of the short-read mappers of the late 2000s, most importantly **Bowtie** (Langmead et al. 2009) and **BWA** (Li and Durbin 2009). Reads are not exact copies of the reference: they contain sequencing errors and genetic variants. These mappers therefore extend backward search with *backtracking*: when the interval becomes empty, they try substituting a different base at a recent position, within a limited budget of mismatches and gaps, and use quality values to prioritise the least confident bases. Later tools (BWA-SW (Li and Durbin 2010) and BWA-MEM) use the FM-index to find maximal exact matches as seeds, which are then extended by dynamic programming. This is the same seed-and-extend idea as BLAST, with exact seeds found in an index of the reference. For long, error-prone reads, minimap2 instead uses *minimizers*: for every window of consecutive k-mers it stores only the smallest k-mer according to some ordering, which gives a compact hash-based index. Hits are chained by dynamic programming before the final alignment (Li 2018). Read mapping, mapping quality and variant calling are discussed in Chapter 5.

1.4.7 Choosing a method

Task	Typical tools	Key idea
Find homologues of one or a few proteins	BLAST (blastp), FASTA/SSEARCH	Neighbourhood words, two hits, X-drop gapped extension; Karlin–Altschul E-values
Search millions of queries or metagenomes	DIAMOND, MMseqs2	Reduced alphabets, spaced or similar k-mers, double indexing, vectorised alignment
Find remote homologues	PSI-BLAST, MMseqs2 profiles, HMMER, HHblits	Profiles and profile HMMs (Section 1.6)
Cluster huge sequence sets	Linclust, MMseqs2, CD-HIT	Compare each sequence only with a few cluster centres
Search large structure databases	Foldseek	3Di structural alphabet with sequence-search (MMseqs2) machinery
Accurate pairwise structural alignment	DALI, TM-align	DALI: compares the intramolecular distance patterns of the two structures; TM-align: superposition and alignment that optimise the TM-score

Task	Typical tools	Key idea
Map reads to a reference genome	Bowtie, BWA, minimap2	Index the reference: FM-index (short reads), minimizers (long reads)
Align mRNAs to a genome	BLAT, minimap2	Genome *k*-mer index held in memory

In all cases, the same trade-off applies. A heuristic is defined by its filter, and every filter that makes a search faster can also make it miss something. Understanding *what* a filter requires, such as two similar 3-mers within 40 residues on one diagonal, tells you *which* alignments can be missed: short ones, ones with many gaps, and ones with no conserved stretch.

1.4.8 From a search hit to an annotation

A list of hits sorted by E-value is not yet a biological conclusion. The step from "this sequence is similar to that one" to "this protein is an X" is where most annotation errors are made (Section 2.2), and it is a favourite topic in oral exams.

Check the query first. The query is itself the result of an experiment and a pipeline: a sample, a sequencing run, an assembly and a gene model. A frameshift, a wrong exon boundary, a truncated open reading frame, a chimeric contig or a contaminant changes what is being aligned. Before searching, it is worth predicting low-complexity regions, transmembrane helices, signal peptides and domain boundaries, and asking whether the sequence is complete.

Read the alignment, not only the E-value. For each hit of interest, look at:

- the bit score and E-value, remembering that the E-value depends on the database searched;
- the percent identity *together with its denominator*. 90% identity over 20 residues says something different from 40% over 600 residues;
- the coverage of *both* the query and the target, and *where* the alignment lies. A tiny E-value over one 80-residue domain says nothing about the other 900 residues of two large proteins;
- whether the alignment is driven by a low-complexity region, a coiled coil, a transmembrane segment or a repeat. Comparing masked with unmasked searches shows how much of the score comes from such regions;
- the domain architecture of the query and the hit (Section 1.9).

The best hit is not the orthologue. The ranking of hits depends on how fast each gene has evolved and on what happens to be in the database. A fast-evolving orthologue can score lower than a conserved paralogue, so the top hit, or even a reciprocal best hit, is only screening evidence. Orthology needs a gene tree (Section 3.4, Chapter 4).

Escalate the method, not the conclusion. If a fast search gives no clear answer, the next step depends on the data: an iterative profile search or a curated profile HMM when a family alignment exists (Section 1.6), a structure search with Foldseek when a reliable (experimental or confidently predicted) structure exists, or an embedding search to suggest candidates (Section 1.10). Each of these produces candidates that must pass the same checks as a BLAST hit (Figure 1.5). Structural similarity can also arise by convergence, and an embedding distance is not a calibrated test of homology.

State a conclusion that matches the evidence. Each step from similarity to homology, from homology to orthology, and from orthology to the same detailed function needs *more* evidence, not a smaller E-value (Chapter 10). Consider a 620-residue bacterial protein whose best hit ($E = 10^{-35}$, 41% identity) is a eukaryotic enzyme, but where the alignment covers only a central domain of about 230 residues,

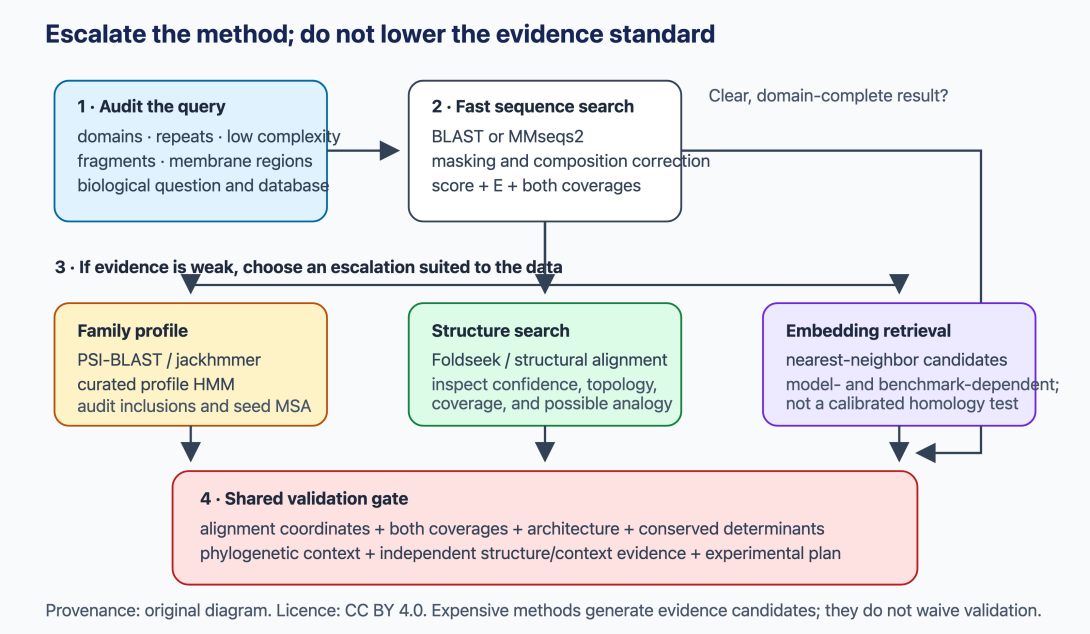


Figure 1.5: A workflow for difficult searches. The query is checked first, and a fast sequence search is run. If the evidence is weak, the search is escalated to profiles, structure or embeddings, depending on the data available. Every candidate, however it was found, then passes the same checks: coordinates and coverage of the alignment, domain architecture, conserved functional residues, phylogenetic context and independent evidence. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

while the N-terminus is low-complexity and the C-terminus has a transmembrane helix. A defensible annotation is "contains a domain of family F", and "putative" enzyme only if the catalytic residues are conserved. Copying the name of the top hit to the whole protein is how errors start to propagate through the databases.

Record what was done. The database and its release, the program and its version, and all parameters (matrix, gap costs, masking, E-value and coverage thresholds, and for iterative searches the inclusion threshold of each round) are part of the result. Without them a search cannot be repeated, and its E-values cannot be compared with those of another search.

See also: further explanation in Section D.1.3; related Wikipedia articles in Section E.1.3.

Sources

Parts of this section are adapted from the Wikipedia articles BLAST (biotechnology), FASTA, Burrows–Wheeler transform, FM-index, Suffix array and Gumbel distribution (CC BY-SA 4.0, retrieved 2 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Sequence Alignment and Search Foundations" (CC BY 4.0), used with permission.

1.5 Alignment-free sequence comparison

Every method so far starts from an alignment, or at least from seeds that are extended into one. Alignment has two drawbacks for some problems. First, it is computationally expensive for very long sequences such as whole genomes. Second, it assumes that the two sequences are *collinear*: homologous

positions appear in the same order. That is not true for genomes that have undergone rearrangements, recombination or horizontal gene transfer. **Alignment-free** methods compare sequences through summary statistics, most often the sets or frequencies of their short words, without computing an alignment at all (Zielezinski et al. 2017).

1.5.1 Word (k-mer) statistics

The simplest approach represents each sequence by the counts of all its k-mers, the words of length k. There are 4^k possible k-mers for DNA and 20^k for proteins. Two sequences are then compared by a distance between their count vectors, for example the Euclidean distance, or the D_2 statistic, which counts the number of shared k-mer occurrences. Variants correct the counts for what is expected from the composition of shorter words, so that only *unexpected* word frequencies contribute. The choice of k matters. A small k gives words that occur by chance in any sequence. A large k gives words that are almost never shared once the sequences have diverged. Other families of methods use the lengths of the longest common substrings, matches of spaced words (as in the spaced seeds of Section 1.4), or information-theoretic measures such as how well one sequence compresses given the other.

1.5.2 MinHash and Mash

For large-scale comparisons the most successful idea comes from computer science. The similarity of two sets A and B, for example the sets of k-mers of two genomes, can be measured by the **Jaccard index** $J = |A \cap B| / |A \cup B|$. Computing it exactly requires storing all k-mers. **MinHash** (Broder 1997) estimates it from a small *sketch*. Every k-mer is mapped to a number by a hash function, and only the s smallest hash values of each set are kept, typically $s = 1000$. The fraction of shared values among the smallest s values of the union is an unbiased estimate of J. A genome of millions of bases is thus summarised in a few kilobytes, and two sketches are compared in microseconds.

Mash (Ondov et al. 2016) turned this into a tool for genomics. It also converts the Jaccard index into an estimate of the evolutionary distance. Under a simple model in which mutations are independent and each one destroys the k-mers that overlap it, the Mash distance is

$$D = -\frac{1}{k} \ln \frac{2J}{1 + J},$$

which approximates the proportion of differing positions (substitutions per site) for closely related sequences. Mash can compare a new genome against all sequenced genomes in seconds, cluster tens of thousands of genomes, or find which known organisms are present in a metagenome.

The same principle of selecting a small, consistent subset of k-mers appears in other tools of this chapter. Minimizers in minimap2 (Section 1.4) and the k-mer selection in Linclust are both ways of finding candidate matches without looking at every word.

1.5.3 Applications and limitations

Alignment-free methods are now standard where speed matters more than detail:

- **Species delimitation and genome comparison.** The *average nucleotide identity* (ANI) between two prokaryotic genomes, computed rapidly from matching fragments by tools such as FastANI, shows a clear gap at about 95%, which is widely used as an operational species boundary for bacteria and archaea, although exceptions occur (C. Jain et al. 2018).

- **Metagenomic classification.** Kraken (Wood and Salzberg 2014) assigns each read to a taxon by looking up its k-mers in a database that maps every k-mer to the lowest common ancestor of all genomes that contain it. Millions of reads can be classified per minute.
- **Transcript quantification.** Pseudo-alignment in tools such as kallisto determines which transcripts a read is compatible with from shared k-mers, without aligning it base by base (Chapter 7).
- **Phylogenies of many genomes**, where the distances are used as input to distance methods such as neighbour joining (Chapter 4).

These measures answer related but different questions, and they should not be confused. The *Jaccard index* of k-mer sets (estimated by MinHash) measures how much sequence *content* two sets share. It depends on k and is lowered by differences in genome size or gene content even when the shared parts are identical. *ANI* measures the identity of the *shared* part only: it is computed over fragments that can be matched, and is reported together with the fraction of each genome that was matched. The *Mash distance* converts a Jaccard index into an approximate substitution rate, but only under its simple mutation model and only for close relatives. *Homology*, finally, is a statement about common ancestry (1). A low Jaccard index or the absence of shared k-mers is not evidence against homology; it usually means only that the sequences have diverged beyond what exact words can detect.

The limitations follow from the method. Exact k-mer matches disappear quickly as sequences diverge, so alignment-free methods are accurate for closely related sequences, for example within a species or a genus. They are poor at detecting distant homology, where alignment with substitution matrices and profiles is far more sensitive. They also give no alignment, so they cannot say *which* positions correspond or where the differences are. In practice, the two approaches are complementary. Alignment-free methods are used to decide quickly which sequences to compare, and alignment is used to analyse the comparisons that matter.

See also: further explanation in Section D.1.4; related Wikipedia articles in Section E.1.4.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Alignment-free sequence analysis and MinHash (CC BY-SA 4.0, retrieved 2 October 2026).

1.6 Profiles and multiple sequence alignments

Pairwise methods compare one sequence with another. However, homologous sequences rarely come in pairs. A protein usually belongs to a *family* of tens to millions of related sequences, and the family as a whole carries much more information than any single member. When many homologues are aligned, it becomes visible which positions are invariant (often catalytic or structurally essential residues), which tolerate only certain kinds of amino acid (e.g. only hydrophobic ones in a buried strand), and where insertions and deletions are common (usually surface loops). This section covers how such alignments are built, how short conserved signals (motifs) are described and found, and how the information in an alignment is condensed into a *profile* that can search databases far more sensitively than a single sequence. The probabilistic version of the profile, the profile hidden Markov model, is described in the next part of this section.

1.6.1 Multiple sequence alignment

1.6.1.1 What a multiple alignment is

A multiple sequence alignment (MSA) of N sequences $S_1, \dots, S_N$ is obtained by inserting gaps into each sequence so that all of them have the same length $L \geq \max_i |S_i|$. The aligned sequences are written as rows of a matrix, and no column may consist only of gaps. Removing the gaps from a row gives back the original sequence. The intended meaning of a column is that its residues are *homologous*: they descend from the same position in a common ancestral sequence (1). In practice, a column is a hypothesis of homology. It is most reliable in conserved core regions and least reliable in gap-rich, variable loops.

Multiple alignments are among the most widely used objects in bioinformatics. They are the input to phylogenetic tree reconstruction (Chapter 4). They are used to identify conserved functional residues and sequence motifs, to build profiles and profile HMMs for database search (see the rest of this section and Section 1.9), and to define protein families. They are also central to structure prediction: the coevolution signal that made modern contact and structure prediction possible, including AlphaFold, is extracted from deep MSAs (Chapter 9).

1.6.1.2 Scoring a multiple alignment

To speak of a "best" alignment, we need an objective function. The most common one is the **sum-of-pairs (SP) score**. Each column is scored by adding the substitution score for every pair of residues in it, and the alignment score is the sum over columns minus gap penalties:

$$\mathrm{SP} = \sum_{\text{columns } c} \sum_{i<j} s(a_{ic}, a_{jc}),$$

where a_{ic} is the residue (or gap) of sequence i in column c and s is taken from a substitution matrix such as BLOSUM62 (Section 1.3). Gap–gap pairs are usually scored as zero. Equivalently, the SP score is the sum of the scores of all $N(N-1)/2$ pairwise alignments *induced* by the MSA.

The SP score is simple, but it has known weaknesses. It implicitly treats every sequence as equally independent, so a family dominated by many near-identical sequences gives them too much weight. This is why practical methods use *sequence weights* (see below). It also lacks a sound evolutionary interpretation: in a column with N copies of the same residue the score grows as N^2, although the evidence comes from a single ancestral residue that has been conserved. Alternative column scores include **entropy-based scores**, where a column c with residue frequencies $p_{a,c}$ is scored by $H_c = -\sum_a p_{a,c} \log p_{a,c}$ and a good alignment *minimises* the total entropy. There are also tree-based scores that count the substitutions required on a phylogeny. When alignment programs are compared with reference alignments, two further measures are used. The *SP accuracy* is the fraction of correctly aligned residue pairs, and the *total column (TC) score* is the fraction of reference columns that are reproduced exactly.

It is worth stating plainly that there is **no single biologically correct objective function** for a multiple alignment. The true alignment is defined by evolutionary history (which residues descend from the same ancestral position), and that history is not observed. Every practical method therefore optimises a *surrogate*: the SP score (progressive methods such as ClustalW), agreement with a library of pairwise alignments (*consistency*, T-Coffee and the accurate MAFFT modes), expected accuracy under a probabilistic model of alignment (ProbCons, HMM-based aligners), a phylogeny-aware treatment that distinguishes insertions from deletions on a guide tree (PRANK), or agreement with known or predicted structures (structure-aware methods such as 3D-Coffee). These surrogates can disagree, and

the "best" alignment under one need not be the best under another. Benchmarks, which use structural superposition as a proxy for homology, are themselves a further surrogate.

1.6.1.3 Exact alignment is intractable

Dynamic programming (Section 1.3) extends directly from two to N sequences. Instead of a two-dimensional matrix we fill an N-dimensional lattice with one axis per sequence. Each cell can be reached from $2^N - 1$ predecessor cells (every non-empty subset of the sequences advances by one residue). For sequences of length L this costs $O(L^N)$ memory and $O(2^N L^N)$ time. Already three proteins of length 300 need 2.7×10^7 cells; ten proteins of that length would need $\approx 6 \times 10^{24}$.

Two observations made exact methods somewhat more practical. First, Carrillo and Lipman showed in 1988 (Carrillo and Lipman 1988) that the score of each induced pairwise alignment can be bounded using the optimal pairwise alignments. The search can therefore be restricted to the region of the lattice near the intersection of the pairwise optima. The program MSA (Lipman et al. 1989) used this idea to align up to about eight moderately long sequences optimally under the SP score. Second, and more fundamentally, the problem cannot be solved efficiently in general: Wang and Jiang (1994) proved that finding the optimal SP alignment is NP-hard. All widely used MSA programs are therefore *heuristics*. They give good alignments in practice, but there is no guarantee that they find the optimum of any objective function.

1.6.1.4 Progressive alignment

The heuristic behind almost all practical MSA programs is **progressive alignment**, introduced by Feng and Doolittle (1987). The idea is to follow the evolutionary history. Sequences that are closely related can be aligned reliably, so they are aligned first, and the growing alignments are then combined in the order given by an approximate tree. The method has three steps:

1. **Pairwise distances.** All $N(N-1)/2$ pairs are compared, either by full dynamic programming or, for speed, by counting shared short words (k-tuples). The similarity of each pair is converted into a distance.
2. **Guide tree.** A tree is built from the distance matrix with a fast clustering method, typically UPGMA or neighbour joining (Chapter 4). The guide tree is not intended as a phylogeny. It only determines the order in which sequences are joined.
3. **Progressive joining.** Following the guide tree from the leaves to the root, sequences and sub-alignments are aligned to each other: sequence to sequence, sequence to alignment, or alignment to alignment. Aligning two alignments ("profile–profile alignment") is ordinary pairwise dynamic programming in which each "residue" is a column, scored for example by the average substitution score over all residue pairs between the two columns.

Progressive alignment needs only $O(N^2)$ pairwise comparisons plus $N-1$ profile alignments, so it scales to thousands of sequences. Its fundamental weakness is greed. Feng and Doolittle stated it as **"once a gap, always a gap"**: decisions taken early, when two close sequences are aligned, are frozen into the sub-alignment and are never revised when more distant sequences are added. In their method, gaps placed early were replaced by a neutral symbol so that they would not be penalised again. Errors in the early alignments, or in the guide tree, therefore propagate to the final result. This matters most when all sequences are distantly related, because then no reliable early alignments exist.

1.6.1.5 The Clustal family

The Clustal programs, developed by Des Higgins and colleagues, have been the most widely used implementations of progressive alignment, and two Clustal papers are among the most cited scientific papers of all time.

- **Clustal (1988).** The first program (Higgins and Sharp 1988) was written to run on the personal computers of the time. Its name reflects the use of *clustering* (UPGMA) to build the guide tree, computed from fast k-tuple similarity scores. ClustalV (1992) was a rewrite in C. It added profile alignment (aligning an alignment to an alignment) and neighbour-joining trees.
- **ClustalW (1994).** The W stands for *weighting*. Thompson et al. (1994) added three improvements that made progressive alignment much more accurate and that remain standard ideas:

  - *Sequence weighting.* Weights are derived from the branch lengths of the guide tree, so that groups of near-identical sequences are down-weighted and divergent sequences are up-weighted.
  - *Varying substitution matrices.* The choice of matrix depends on the divergence of the sequences being aligned. Closely related sequences are aligned with "hard" matrices (high-identity BLOSUM or PAM matrices), divergent ones with "soft" ones.
  - *Position-specific gap penalties.* Gap-opening penalties are reduced in runs of hydrophilic residues, which usually form surface loops, and at positions where the existing alignment already has gaps. They are increased close to existing gaps, so that gaps cluster in the same places instead of being scattered.

  In addition, the joining of very divergent sequences is delayed until the rest of the alignment has been built. ClustalW uses neighbour joining for the guide tree. Because all pairwise distances are computed, the method is $O(N^2)$ in the number of sequences. ClustalX (1997) added a graphical interface, and Clustal 2 (2007) packaged both.
- **Clustal Omega (2011).** By the 2000s, families with tens or hundreds of thousands of members had become common, and computing all pairwise distances was no longer feasible. Clustal Omega (Sievers et al. 2011) replaced the core of the method:

  - Guide trees are computed with **mBed**. Each sequence is represented by its distances to a small set of "seed" sequences (an embedding in a low-dimensional space). The sequences are clustered (k-means) in that space, and a UPGMA tree is built within clusters. This reduces the cost from $O(N^2)$ to $O(N \log N)$.
  - The progressive joining step uses **HHalign** from the HH-suite. Each sub-alignment is turned into a profile hidden Markov model, and the two HMMs are aligned, which is more sensitive than simple profile–profile scoring.

  As a result, Clustal Omega can align more than 100,000 sequences on a single processor in a few hours, while keeping accuracy comparable to the best methods on small benchmarks. Its authors also stressed that alignment accuracy *decreases* when very many sequences are aligned with most methods, and that big alignments need methods designed for them.

1.6.1.6 Consistency and iterative refinement

Two further ideas address the greediness of progressive alignment.

Consistency. If residue x in sequence A aligns with residue y in sequence C, and y aligns with residue z in B, then the pairwise alignment of A and B should preferably align x with z. T-Coffee (Notredame et al. 2000) uses this transitivity. It first computes a *primary library* of pairwise alignments of all pairs, using both global alignments (from ClustalW) and local ones (from LALIGN). Each aligned residue pair is

weighted by the identity of the alignment it came from. The library is then *extended*: the weight of a pair (x, z) is increased by the support from every third sequence C through which x and z are consistently aligned. Finally, an ordinary progressive alignment is carried out, but using the extended library weights instead of a substitution matrix as the scores. Information from all sequences thus influences every pairwise step, and errors made early are much less likely. The cost is time and memory that grow as $O(N^3 L^2)$, which limits T-Coffee to at most a few hundred sequences. Consistency is also the basis of probabilistic aligners such as ProbCons, and of the most accurate modes of MAFFT.

Iterative refinement. An alignment can be improved after it has been built. A typical refinement step divides the sequences into two groups (for example by deleting one edge of the guide tree), realigns the two sub-alignments to each other, and keeps the result if the objective score improves. **MUSCLE** (Edgar 2004) combines progressive alignment with such refinement in three stages:

1. A *draft progressive* alignment using fast k-mer distances and a UPGMA tree.
2. An *improved progressive* alignment: the tree is rebuilt from the (Kimura-corrected) identities of the draft alignment, and the alignment is recomputed for the parts of the tree that changed.
3. *Refinement* by tree-dependent restricted partitioning: edges of the tree are deleted in turn, the two resulting profiles are realigned, and the change is accepted if the score improves.

MUSCLE scores profile columns with a "log-expectation" function, and it was notable for being both faster and more accurate than ClustalW and T-Coffee on the benchmarks of the time.

MAFFT, first published by Kazutaka Katoh and colleagues in 2002 (Katoh et al. 2002; Katoh and Standley 2013), introduced a different speed-up for the pairwise step. Each amino acid is represented by numerical values of its volume and polarity. The correlation between two sequences at every possible relative offset can then be computed with the **fast Fourier transform (FFT)** in $O(L \log L)$ time. Peaks in this correlation reveal homologous segments, and dynamic programming is restricted to the regions between them. MAFFT offers a range of strategies. FFT-NS-2 is a fast progressive method. FFT-NS-i adds iterative refinement. L-INS-i, G-INS-i and E-INS-i are consistency-based iterative methods using local, global or generalised pairwise alignments, and they are among the most accurate methods available for up to a few hundred sequences. Modern versions also include options for very large data sets and for adding new sequences to an existing alignment.

1.6.1.7 How good are multiple alignments?

Because the true alignment is unknown, MSA methods are evaluated on reference alignments in which the correct homologous positions are believed to be known. The most widely used benchmarks, such as BAliBASE (Thompson et al. 1999), derive the reference from superpositions of known three-dimensional structures, since structurally equivalent residues are the best available proxy for homologous ones. Methods are compared by the SP accuracy and the TC score defined above. Three conclusions have held for decades:

- For sequences above about 30–40% identity, all good methods agree and are essentially correct. Below about 20–25% identity (the "twilight zone",

    1) all methods make many errors, and the differences between them become large.

- Consistency-based and iterative methods are the most accurate on small and medium-sized sets. Clustal Omega and the fast modes of MAFFT are the practical choices for very large sets.
- Alignment errors are concentrated in loops and at the ends of sequences. They can affect downstream analyses, for example phylogenetic trees (Chapter 4), so variable regions are often removed ("trimmed") before tree building. Phylogeny-aware aligners such as PRANK (Löytynoja and Goldman 2008) treat insertions and deletions more realistically than methods that systematically minimise the number of gaps.

1.6.1.8 Alignment uncertainty

A multiple alignment is an estimate, but it is usually used as if it were data. A tree, a conservation score or a profile built from it inherits its errors, and nothing in a single alignment shows which columns are reliable. Different programs, different guide trees and even different input orders can give different alignments of the same sequences, and the differences are concentrated in exactly the divergent regions that matter most for remote homology and deep phylogeny.

One response is to make the uncertainty visible. **MUSCLE5** (Edgar 2022) builds an *ensemble* of alignments by perturbing the guide tree and the scoring parameters. The fraction of ensemble members in which a column, or a whole downstream result such as a branch of a tree, is recovered is a measure of its confidence. Its benchmarks also showed that a single confident-looking alignment can hide considerable uncertainty, and that conclusions which change between equally plausible alignments should not be trusted. Consistency-based aligners such as T-Coffee and probabilistic aligners can likewise report a reliability score for each column.

Trimming (removing unreliable columns) is a related choice. It can reduce noise, but it also removes fast-evolving positions that carry real signal (Section 4.4). Whether and how to trim is therefore a modelling decision, best tested by checking whether the conclusion changes.

1.6.2 Sequence motifs

A **sequence motif** is a short, recurring pattern in DNA, RNA or protein sequences that is assumed to have a biological function. Examples are transcription factor binding sites in DNA, splice signals in RNA, and in proteins, catalytic sites, metal-binding sites, sorting signals and sites of post-translational modification. Motifs are often only 5–20 residues long. They may lie in otherwise unrelated sequences: regulatory sites in promoters are not, in general, homologous to each other. Such sites may have arisen independently many times, because only the binding specificity of the protein recognising them is conserved (1). Short protein motifs such as phosphorylation sites can also arise by convergence.

1.6.2.1 Consensus sequences and patterns

The simplest description of a motif is a **consensus sequence**: the most common residue at each position, for example `TATAAT` for the bacterial promoter −10 box (the "Pribnow box"). A consensus discards the information about how variable each position is. Very few real sites actually match the consensus exactly.

A more expressive description is a **pattern**, written as a regular expression. The PROSITE database, created by Amos Bairoch in 1988 (Bairoch 1991; Sigrist et al. 2012), made such patterns standard for proteins. In PROSITE notation:

- elements are separated by `-`;
- `x` means any residue;
- `[ST]` means S or T;
- `{P}` means any residue except P;
- `x(2,4)` means between two and four arbitrary residues;
- `<` and `>` anchor the pattern to the N- or C-terminus.

Three well-known examples:

Motif	PROSITE pattern
N-glycosylation site	`N-{P}-[ST]-{P}`
ATP/GTP-binding P-loop (Walker A)	`[AG]-x(4)-G-K-[ST]`
C2H2 zinc finger	`C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H`

Patterns are easy to read and to search for. However, matching is binary: a sequence either matches or it does not. Every allowed residue counts the same, however rare it is at that position, and a single unusual residue causes a miss. Short patterns also match many sequences by chance; the N-glycosylation pattern is found in most proteins, while only a minority of the matches are real glycosylation sites. PROSITE therefore documents the expected numbers of true and false matches for each pattern. For divergent families it uses *profiles* (below) instead.

1.6.2.2 Position weight matrices

A **position weight matrix** (PWM), also called a position-specific scoring matrix (PSSM), assigns a score to every residue at every position of a gap-free motif of length w. It is built from a set of aligned sites in three steps:

1. **Counts.** $n_{b,j}$ is the number of sites with residue b at position j (the *position frequency matrix*).
2. **Probabilities.** These are estimated with **pseudocounts**, which prevent residues that are absent from a small training set from getting probability zero (and a score of $-\infty$):

$$p_{b,j} = \frac{n_{b,j} + \alpha_b}{N + \sum_{b'} \alpha_{b'}},$$

   where N is the number of sites and α_b the pseudocount for residue b.
3. **Log-odds weights.** The probabilities are compared with a background model q_b, typically the overall base or amino acid composition:

$$W_{b,j} = \log_2 \frac{p_{b,j}}{q_b}.$$

The score of a candidate site $x_1 \ldots x_w$ is the sum of the weights, $\sum_j W_{x_j,j}$. It is the log-likelihood ratio between "this is an instance of the motif" and "this is background sequence", under the assumption that positions are independent. A score above zero means the site is more likely under the motif model than under the background. In practice, a threshold is set to balance true and false matches, as for the substitution scores of Section 1.3, which are log-odds scores of the same kind. For DNA-binding proteins, the PWM score is also approximately proportional to the binding energy, which gives the model a physical interpretation (Stormo 2000).

Worked example (illustrative data):

Ten aligned sites (illustrative): `TATAAT`, `TATAAT`, `TAAAAT`, `TATACT`, `GATAAT`, `TATGAT`, `TACAAT`, `TATATT`, `AATAAT`, `TATAAG`.

Counts (position frequency matrix):

	1	2	3	4	5	6
A	1	10	1	9	8	0
C	0	0	1	0	1	0
G	1	0	0	1	0	1
T	8	0	8	0	1	9

Log-odds weights $W_{b,j} = \log_2(p_{b,j}/q_b)$ in bits, with a pseudocount of 0.25 per base and a uniform background $q_b = 0.25$:

	1	2	3	4	5	6
A	-1.14	+1.90	-1.14	+1.75	+1.58	-3.46
C	-3.46	-3.46	-1.14	-3.46	-1.14	-3.46
G	-1.14	-3.46	-3.46	-1.14	-3.46	-1.14
T	+1.58	-3.46	+1.58	-3.46	-1.14	+1.75

Information content per position (bits, from the raw frequencies): 1.08, 2.00, 1.08, 1.53, 1.08, 1.53; total 8.3 bits.

Scores of three test sequences (sum of the weights at each position):

Sequence	Score (bits)
`TATAAT (consensus)`	+10.15
`TACGAT`	+4.54
`GCGCGC`	-18.43

The consensus `TATAAT` gets the highest possible score. `TACGAT`, which differs at two positions, still gets a clearly positive score. A GC-rich sequence gets a strongly negative score.

How much a motif stands out from the background is measured by its **information content**. For position j it is the relative entropy (Kullback–Leibler divergence) between the motif and the background distributions:

$$I_j = \sum_b p_{b,j} \log_2 \frac{p_{b,j}}{q_b}.$$

With a uniform background this is $I_j = \log_2 |\mathcal{A}| - H_j$, where H_j is the Shannon entropy of the column and $|\mathcal{A}|$ is the alphabet size. The maximum is therefore 2 bits for DNA and $\log_2 20 \approx 4.3$ bits for proteins. **Sequence logos** (Schneider and Stephens 1990) display this graphically: the height of the stack at each position is I_j, and each letter's height is proportional to its frequency (Figure 1.6). For small samples a correction term is usually subtracted, because entropy estimated from few sequences is biased downwards.

The total information content of a motif also says how often it is expected to occur by chance. A motif with I bits matches a random sequence about once every 2^I positions. The 8.3 bits of the example above therefore correspond to one chance match every few hundred base pairs, which is why real promoters are recognised by combinations of signals, not by the -10 box alone.

PWMs assume that positions contribute independently and that motifs have no insertions or deletions. Models with dependencies between positions exist. Allowing insertions and deletions leads to profile HMMs, described later in this section.

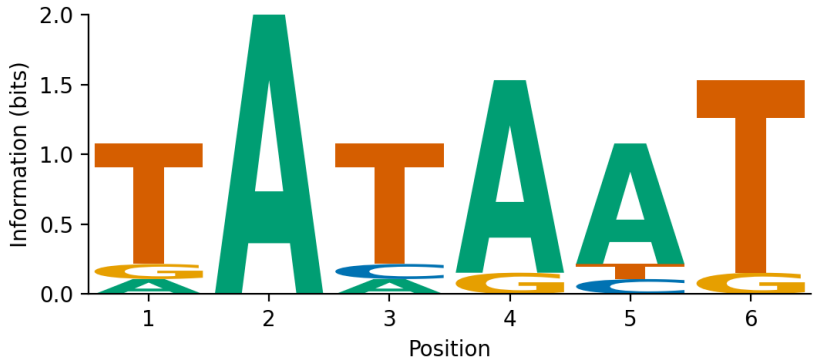


Figure 1.6: Sequence logo of the ten illustrative −10 box sites above. The height of each stack is the information content of the position in bits (uniform background, no small-sample correction). Within a stack, letters are scaled by their frequency.

1.6.2.3 Finding motifs de novo

Often the sites are not known in advance. We may have, for example, a set of promoters of co-regulated genes and want to find the motif that they share. This **motif-finding problem** can be stated as follows. Given N sequences, find one site of width w in each (or in most), such that the sites are as similar to each other as possible, for example as measured by the information content of the resulting PWM. The search space is huge (about L^N combinations of starting positions for sequences of length L), so stochastic and statistical methods are used.

Gibbs sampling (Lawrence et al. 1993) is a Markov chain Monte Carlo method. The algorithm is:

1. Choose a random starting position a_i for a site in each sequence i.
2. Choose one sequence z and remove it.
3. Build a PWM (with pseudocounts) from the current sites in the remaining $N - 1$ sequences.
4. For every possible position x in sequence z, compute the ratio Q_x/P_x. Here Q_x is the probability of the w-long segment starting at x under the PWM, and P_x its probability under the background model.
5. Sample a new site position a_z with probability proportional to Q_x/P_x, i.e. randomly rather than greedily.
6. Repeat from step 2 until the motif stops improving.

Because the new position is sampled rather than chosen as the best, the algorithm can escape from poor local optima, and it gradually converges on a strong shared motif if one exists. Lawrence and colleagues also introduced "phase shift" moves, which shift all sites a few positions left or right, because the sampler otherwise tends to get stuck on a shifted version of the true motif. Gibbs sampling was originally applied to find subtle shared motifs in sets of weakly similar proteins, but it became particularly popular for transcription factor binding sites.

MEME (Multiple EM for Motif Elicitation; Bailey and Elkan (1994)) uses the **expectation–maximisation (EM)** algorithm instead. The data are modelled as a mixture of motif occurrences, described by a PWM, and background sequence. EM alternates two steps:

- **E-step:** for every position in every sequence, compute the posterior probability that a motif occurrence starts there, given the current PWM.
- **M-step:** re-estimate the PWM as a weighted average over all positions, weighted by those probabilities.

Each iteration cannot decrease the likelihood, so EM converges, but only to a local optimum. MEME therefore starts EM from many initial PWMs built from substrings that actually occur in the data, and keeps the best result. After a motif is found, its occurrences are down-weighted and the search is repeated to find further motifs. MEME also made the model of how many sites a sequence contains explicit:

- **OOPS**: exactly one occurrence per sequence;
- **ZOOPS**: zero or one occurrence per sequence;
- **TCM** (two-component mixture): any number of occurrences.

Each motif is reported with an E-value for its significance, and the companion program MAST searches databases with the discovered motifs.

Both approaches have the same basic limitations. They need the motif width in advance (or try several), and they find statistically over-represented patterns, which may include repeats or low-complexity sequence as well as functional sites. A motif found de novo is a hypothesis, which needs support from experiments or from conservation across species.

1.6.3 Profiles: from patterns to position-specific scoring

1.6.3.1 PROSITE patterns versus profiles

The limitations of patterns, together with the observation that the conserved positions of a family are far more informative than the variable ones, led to **profiles**. Gribskov et al. (1987) defined a profile as a position-specific scoring table derived from a multiple alignment. It has one row for each alignment position, with a score for each of the 20 amino acids and a position-specific penalty for opening and extending gaps. A query sequence is aligned to the profile by ordinary dynamic programming, using the profile row instead of a fixed substitution score at each position.

In Gribskov's construction, the score for amino acid b at position j is an average of substitution-matrix scores, weighted by the residues seen in the alignment column: $W_{b,j} = \sum_a f_{a,j}\, s(a,b)$. Here $f_{a,j}$ is the frequency of residue a in column j and s is a substitution matrix such as PAM250. This simple formula already has the crucial property that a residue never seen at a position, but similar to those that are, still scores well. The profile therefore generalises beyond the sequences it was built from.

The differences from a pattern can be summarised as:

	Pattern (regular expression)	Profile
Representation	Allowed residues per position	Score for every residue at every position
Matching	Yes/no	Graded score; significance estimated
Insertions/deletions	Fixed ranges such as `x(2,4)`	Position-specific gap penalties
Rare but allowed residues	Weighted like common ones	Lower score
Best suited for	Short, highly conserved sites	Whole domains and divergent families

PROSITE added such "generalised profiles" in the 1990s to cover families that patterns could not describe (Sigrist et al. 2012). The profile idea is also the basis of PSI-BLAST and, in probabilistic form, of the profile HMMs in Pfam (Section 1.9).

1.6.3.2 Building a good profile

How well a profile works depends on how it is estimated. Three issues recur in all profile methods, including profile HMMs:

- **Sequence weighting.** Sequence databases are very redundant. If a family alignment contains 50 nearly identical sequences from one subfamily and five divergent ones, unweighted frequencies describe that subfamily, not the family. Weighting schemes give less weight to sequences that are similar to many others. A widely used simple scheme is the position-based weighting of Henikoff and Henikoff (Henikoff and Henikoff 1994). In each column with r different residue types, a sequence whose residue occurs s times receives $1/(rs)$, and these contributions are summed over all columns. ClustalW's tree-based weights serve the same purpose.
- **Pseudocounts and priors.** A column with few sequences, or few *effective* sequences after weighting, gives unreliable frequencies. Pseudocounts are added to fill in residues that have not been observed. Better than uniform pseudocounts are ones that use what we know about amino acid substitutions. One option is to derive them from a substitution matrix: residues similar to those observed get more pseudocounts. Another is **Dirichlet mixture priors** (Sjölander et al. 1996), which encode typical kinds of alignment columns ("hydrophobic", "small", "aromatic", ...) learnt from many alignments. The fewer the sequences, the more the estimate relies on the prior.
- **Position-specific gaps.** Insertions and deletions are frequent in loops and rare in the core. Profiles that use lower gap penalties where the training alignment has gaps align new sequences more accurately.

1.6.3.3 A worked PSSM

A small example shows how the pieces fit together (Figure 1.7). Let column i of an alignment contain $n_i(a)$ copies of amino acid a and N_i residues in total (gaps are not counted), and let $p(a)$ be the background frequency of a. With a pseudocount of total weight α, distributed in proportion to the background, the estimated frequency and the score in bits are

$$\hat{q}_i(a) = \frac{n_i(a) + \alpha\, p(a)}{N_i + \alpha}, \qquad w_i(a) = \log_2 \frac{\hat{q}_i(a)}{p(a)}.$$

Take the six aligned sequences `ADKLG`, `ADKLG`, `AEKLG`, `ADQLG`, `-DKLG` and `ADKIG`, and $\alpha = 1$. Column 2 has five D and one E. With $p(\mathrm{D}) = 0.059$, $\hat{q}_2(\mathrm{D}) = (5 + 0.059)/7 = 0.72$ and $w_2(\mathrm{D}) = \log_2(0.72/0.059) = +3.6$ bits. With $p(\mathrm{E}) = 0.058$, E scores +1.4 bits. A residue not seen in the column, such as A, gets $\hat{q}_2(\mathrm{A}) = p(\mathrm{A})/7$ and therefore $\log_2(1/7) = -2.8$ bits: a penalty, but not the $-\infty$ that raw frequencies would give. (With this simple background-weighted pseudocount, every unobserved residue gets the same score; a substitution-matrix or Dirichlet prior would penalise I less than W in a column of L, for example.) The score of an ungapped sequence segment $x_1 \ldots x_L$ is the sum $\sum_i w_i(x_i)$, one cell per column. A positive sum does not by itself mean that the sequence belongs to the family: as for any score, a threshold or an E-value must be calibrated, and gaps need their own model.

1.6.3.4 PSI-BLAST

PSI-BLAST (Position-Specific Iterated BLAST; Altschul et al. (1997)) made profile searching fast, automatic and widely used. It builds a profile from its own search results, without a pre-made family alignment:

1. **Round 1** is an ordinary BLAST search (Section 1.4) of the query against a large database such as UniRef or NCBI nr.

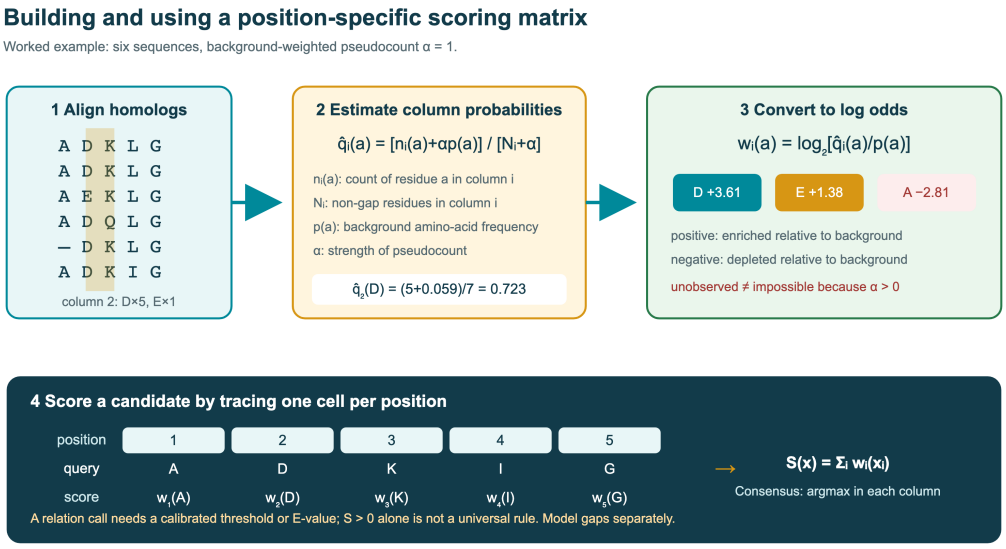


Figure 1.7: Building and using a position-specific scoring matrix for the six-sequence example in the text: (1) count residues in each column of the alignment, (2) estimate column frequencies with background-weighted pseudocounts, (3) convert them to log-odds scores in bits, and (4) score a new sequence by adding one cell per position. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

2. All hits with an E-value below an **inclusion threshold** (typically 0.001–0.005) are collected into a multiple alignment, which is anchored on the query (columns where the query has a gap are ignored).
3. A **PSSM** with one column per query position is computed from this alignment. The sequences are weighted to reduce redundancy, and pseudocounts are derived from the BLOSUM62 matrix. The amount of pseudocount decreases as the number of effectively independent sequences grows.
4. The PSSM replaces the query and the substitution matrix in the next search round. BLAST's word seeding, extension and Karlin–Altschul statistics work unchanged with position-specific scores.
5. Steps 2–4 are repeated for a fixed number of iterations, or until no new sequences are found ("convergence").

Each round, the profile represents the family better than the query alone did, so homologues too distant to be found in the first round can be found in later ones. In a benchmark based on structurally classified domains in SCOP, Park et al. (1998) showed that profile methods (PSI-BLAST and the HMM-based SAM-T98) detected about **three times as many** remote homologues as pairwise methods (BLAST, FASTA and Smith–Waterman) at the same error rate. This was one of the clearest demonstrations of the value of family information, and PSI-BLAST profiles became the standard input to secondary structure, contact and fold prediction methods in the 1990s and 2000s (Chapter 9).

The iteration also has a well-known failure mode, **profile corruption** (also called profile drift). If a single unrelated sequence passes the inclusion threshold, perhaps through a low-complexity region, a coiled coil or a promiscuous domain shared with the query, it is included in the profile. In the next round it pulls in its own relatives, and after a few iterations the profile may describe a different family. The E-values of the later rounds are then misleading, because they measure similarity to the corrupted profile, not to the original query. The same problem affects every iterative profile search, including jackhmmer, HHblits and MMseqs2 profile searches.

Corruption can often be recognised by inspecting the rounds rather than only the final list. Warning

signs are a sudden jump in the number of hits from one round to the next; new hits that align only to a short part of the query, such as a single low-complexity segment, coiled coil or repeat; and hits whose annotations belong to a clearly different family. Several things reduce the risk:

- a strict inclusion threshold;
- masking low-complexity regions;
- composition-based statistics, which correct scores for sequences with unusual amino acid composition (Schäffer et al. 2001);
- checking that new hits are plausible;
- limiting the number of iterations.

A second limitation is that PSI-BLAST, like BLAST, uses ad hoc gap penalties and treats the profile as a series of independent columns. The probabilistic framework of hidden Markov models, which models insertions and deletions position by position and calculates significance in a principled way, addresses both of these points. It is the subject of the next part of this section.

See also: further explanation in Section D.1.5; related Wikipedia articles in Section E.1.5.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Multiple sequence alignment, Clustal, Sequence motif, Position weight matrix, Sequence logo, Gibbs sampling, Expectation–maximization algorithm, PROSITE and BLAST (biotechnology) (CC BY-SA 4.0, retrieved 2 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Sequence Alignment and Search Foundations" (CC BY 4.0), used with permission.

1.6.4 Hidden Markov models

Position-specific scoring matrices and PSI-BLAST profiles score each column of a family independently, and they treat insertions and deletions with ad hoc, position-independent gap penalties. Hidden Markov models (HMMs) put the whole idea of a profile on a probabilistic footing. Substitutions, insertions and deletions all become probabilities that can be estimated from data, and an alignment of a sequence to a family becomes a path through a probabilistic model. The same small set of dynamic programming algorithms then gives the best alignment, the total probability of the sequence, and the confidence of every aligned residue. HMMs are the basis of HMMER, Pfam and the HH-suite. They are also used, under other names, for gene finding (Chapter 5) and membrane-protein topology prediction (Chapter 9). They are the most important probabilistic model in sequence bioinformatics. The textbook treatment is Durbin et al. (1998); short introductions are Eddy (2004a) and Rabiner (1989).

1.6.4.1 From Markov chains to hidden Markov models

A **Markov chain** is a sequence of random variables $x_1, x_2, \dots, x_L$ in which the probability of each symbol depends only on the previous one:

$$P(x) = P(x_1) \prod_{i=2}^{L} P(x_i \mid x_{i-1}) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}.$$

The transition probabilities a_{st} form a matrix whose rows sum to one. A first-order Markov chain over the four nucleotides can, for example, capture the fact that CG dinucleotides are rare in most of the human genome (the C is often methylated and mutates to T). It can also capture that they are more common in *CpG islands* near promoters. Two such chains, one trained on islands and one on background,

can be used to classify a given stretch of DNA by the log-odds ratio of the two models (Durbin et al. 1998).

Classification of a *given* stretch is not the real problem, however. In a chromosome we do not know where the islands start and end. A **hidden Markov model** solves this by separating what we observe from what we want to infer. An HMM has:

- a set of hidden **states** $k = 1, \dots, N$ (for example "island" and "background", or "exon", "intron" and "intergenic");
- **transition probabilities** $a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ between states, where $\pi = \pi_1 \dots \pi_L$ is the *state path*. The path itself is a Markov chain;
- **emission probabilities** $e_k(b) = P(x_i = b \mid \pi_i = k)$, the probability of observing symbol b when the model is in state k;
- start (and optionally end) probabilities, often written as transitions a_{0k} from a silent begin state 0.

The observed sequence is generated by walking along the state path and emitting one symbol from each state. The joint probability of a sequence x and a path π is simply the product of all transitions and emissions used:

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i)\, a_{\pi_i \pi_{i+1}},$$

with $a_{\pi_L \pi_{L+1}}$ the transition to the end state (or 1 if there is no end state). The model is "hidden" because many different paths can generate the same observed sequence, and in general we only see x.

The standard teaching example, from Durbin et al. (1998), is the **occasionally dishonest casino**. A casino uses a fair die most of the time, but occasionally switches to a loaded die that shows a six half of the time. We observe only the rolls. The hidden states are *Fair* (F) and *Loaded* (L), with emission probabilities $e_F(b) = 1/6$ for all faces, $e_L(6) = 0.5$ and $e_L(b) = 0.1$ for $b = 1, \dots, 5$. The transition probabilities are $a_{FL} = 0.05$ and $a_{LF} = 0.1$. The model is used as a worked example below.

1.6.4.2 History

Hidden Markov models were described by Leonard Baum and colleagues at the Institute for Defense Analyses in Princeton in a series of statistical papers from 1966 to the early 1970s. These included the estimation algorithm now called the **Baum–Welch algorithm**, developed by Baum and Lloyd Welch (Baum et al. 1970). Independently, Andrew Viterbi published in 1967 the dynamic-programming decoding algorithm for convolutional codes in noisy communication channels that now carries his name (Viterbi 1967). The Viterbi algorithm has a history of multiple invention. It is essentially the same recursion that Needleman and Wunsch used for alignment (Section 1.3), and at least seven independent discoveries have been documented.

The first major application was **speech recognition**, from James Baker's work in 1975 and Frederick Jelinek's group at IBM, where the hidden states are phonemes or words and the observations are acoustic features. Lawrence Rabiner's 1989 tutorial (Rabiner 1989) became the standard introduction and brought the methods to a wide audience. In the second half of the 1980s HMMs began to be used to segment DNA, for example by base composition (Churchill 1989). In the early 1990s David Haussler's group in Santa Cruz, with Anders Krogh, introduced **profile HMMs** for protein families (Krogh et al. 1994). Within a few years HMMs were used for gene finding (GENSCAN, GeneMark.hmm; Burge and Karlin (1997); Lukashin and Borodovsky (1998)), transmembrane topology prediction (TMHMM; Krogh et al. (2001)), signal peptides and many other problems. Sean Eddy's 1998 review of profile HMMs (Eddy 1998) and the textbook by Durbin, Eddy, Krogh and Mitchison (Durbin et al. 1998) made them standard bioinformatics tools.

1.6.4.3 The three (four) problems

Rabiner formulated the use of an HMM as three basic problems. Biology adds a fourth, posterior decoding, which turns out to be very important.

Problem	Question	Algorithm
Evaluation	What is the probability $P(x)$ of the sequence, summed over all paths?	Forward
Decoding	Which single path π^* is most probable?	Viterbi
Posterior decoding	What is the probability that residue i was emitted by state k?	Forward–backward
Learning	Which parameters $a_{kl}, e_k(b)$ make the training data most probable?	Baum–Welch (EM), Viterbi training

The naive solution to the first two problems is to enumerate all paths, but there are N^L of them. All the algorithms below avoid this by dynamic programming over the sequence positions. They use the fact that, because of the Markov property, the best (or summed) way to reach state l at position $i+1$ only depends on the values for all states at position i.

1.6.4.3.1 Viterbi: the most probable path

Let $v_k(i)$ be the probability of the most probable path that ends in state k after emitting $x_1 \dots x_i$. Then

$$v_l(i+1) = e_l(x_{i+1}) \max_k \big(v_k(i)\, a_{kl}\big), \qquad v_k(1) = a_{0k}\, e_k(x_1).$$

For each cell we store a *traceback pointer* to the k that gave the maximum. The probability of the best path is $P(x, \pi^*) = \max_k v_k(L) a_{k0}$, and the path itself is recovered by following the pointers back from the best final state. This is exactly the structure of the alignment recursions in Section 1.3: a table filled from left to right, a max over predecessors, and traceback.

1.6.4.3.2 Forward: the probability of the sequence

If we replace the maximum by a sum, we obtain the probability of the sequence summed over all paths. Let $f_k(i) = P(x_1 \dots x_i, \pi_i = k)$. Then

$$f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i)\, a_{kl}, \qquad P(x) = \sum_k f_k(L)\, a_{k0}.$$

$P(x)$ is what we want when we ask whether a sequence belongs to a family. It does not depend on choosing one particular alignment, and it is the natural quantity to compare with a null model.

1.6.4.3.3 Backward and posterior decoding

The **backward** variable $b_k(i) = P(x_{i+1} \dots x_L \mid \pi_i = k)$ is computed by the same kind of recursion, from the end of the sequence:

$$b_k(i) = \sum_l a_{kl}\, e_l(x_{i+1})\, b_l(i+1), \qquad b_k(L) = a_{k0}.$$

Combining the two gives the **posterior probability** that residue i was emitted by state k, given the whole sequence:

$$P(\pi_i = k \mid x) = \frac{f_k(i)\, b_k(i)}{P(x)}.$$

Posterior decoding chooses, at each position, the state with the highest posterior probability. It can give a better per-residue accuracy than the Viterbi path, but the resulting sequence of states is not guaranteed to be a valid path (it may use a transition with zero probability). More importantly, the posteriors measure how *certain* each part of the result is. In profile HMM alignments they tell which aligned residues are reliable and which are essentially guesses. HMMER reports them under each alignment.

1.6.4.3.4 Complexity and numerical issues

For an HMM with N states and a sequence of length L, all three algorithms need $O(N^2 L)$ time, because every cell considers every predecessor state, and $O(NL)$ memory. If the model has a sparse transition structure, as profile HMMs do, the time is proportional to the number of non-zero transitions, $O(NL)$ for a profile. The Viterbi path can be found in linear memory with the same divide-and-conquer trick used for pairwise alignment (Hirschberg; Myers and Miller (1988)).

The probabilities quickly become smaller than the smallest number a computer can represent: a product of a few hundred probabilities below 0.1 underflows double precision. Viterbi is therefore computed in **log space**, where products become sums and the maximum is unchanged. The forward and backward algorithms need sums of probabilities. They are computed either with **scaling** (normalising the column at each position and keeping track of the scaling factors, as in Rabiner (1989)) or in log space with the identity $\log(e^a + e^b) = \max(a,b) + \log\left(1 + e^{-|a-b|}\right)$, often implemented with a lookup table.

1.6.4.3.5 Learning the parameters

If the state paths of the training sequences are known (for example, annotated genes, or sequences in a curated multiple alignment), the maximum-likelihood parameters are just normalised counts: $a_{kl} = A_{kl} / \sum_{l'} A_{kl'}$ and $e_k(b) = E_k(b) / \sum_{b'} E_k(b')$, where A and E count the transitions and emissions used. Pseudocounts or priors are added to avoid zero probabilities for events that happen not to be observed.

If the paths are unknown, the **Baum–Welch algorithm** estimates the parameters iteratively. It is a special case of the expectation–maximisation (EM) algorithm. In the E step, the forward and backward variables give the *expected* number of times each transition and emission is used, given the current parameters:

$$A_{kl} = \sum_j \frac{1}{P(x^j)} \sum_i f_k^j(i)\, a_{kl}\, e_l(x^j_{i+1})\, b_l^j(i+1), \qquad E_k(b) = \sum_j \frac{1}{P(x^j)} \sum_{i\,:\,x_i^j = b} f_k^j(i)\, b_k^j(i).$$

In the M step, these expected counts replace the observed counts in the formulas above. Each iteration is guaranteed not to decrease the likelihood of the training data, but the algorithm converges to a *local* maximum, so the result depends on the starting point. A cheaper alternative is **Viterbi training**: align all sequences with the current model, count the transitions and emissions on the Viterbi paths, and re-estimate. It optimises the probability of the best paths rather than the full likelihood, and is often used in practice. HMMs can also be trained discriminatively, or with Bayesian priors on the parameters, for example Dirichlet distributions (see below).

1.6.4.4 Worked example: the dishonest casino

To see the algorithms at work, 300 rolls were simulated from the casino model above, and the Viterbi and forward–backward algorithms were then run on the observed rolls only. Figure 1.8 shows the true (hidden) states, the Viterbi path, and the posterior probability that each roll came from the loaded die.

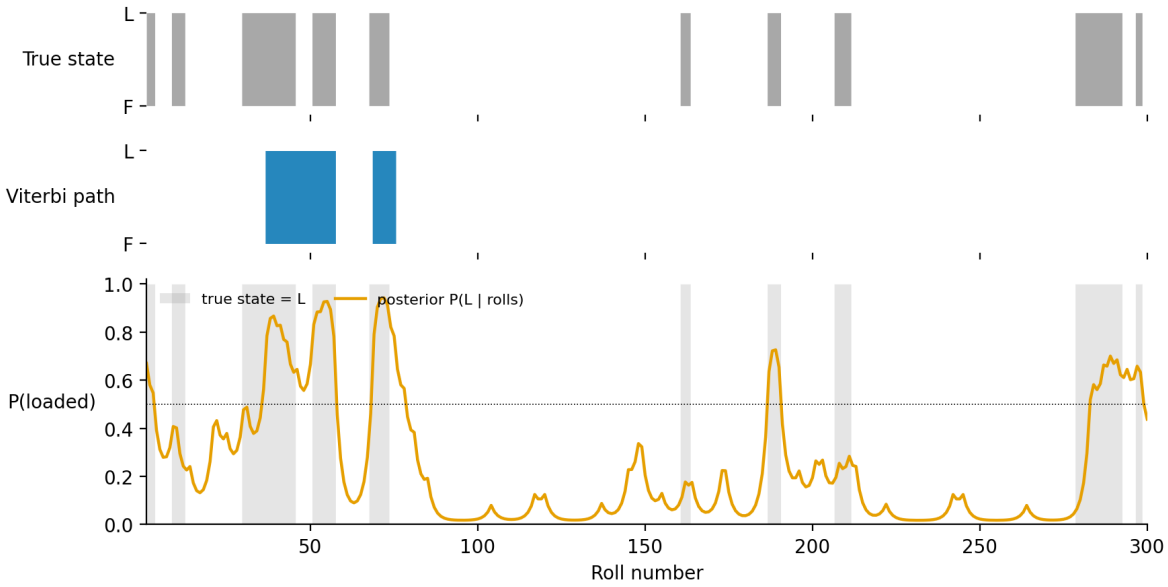


Figure 1.8: The occasionally dishonest casino. Top: the hidden state actually used for each of 300 rolls (grey = loaded). Middle: the Viterbi path decoded from the rolls alone. Bottom: the posterior probability $P(\pi_i = L \mid x)$ from the forward–backward algorithm, with the true loaded segments shaded.

In this run of 300 rolls the loaded die was used for 21% of the rolls and 22% of all rolls were sixes. The Viterbi path assigns the correct state to 83% of the rolls, and thresholding the posterior at 0.5 gives 87%. The log-probability of the single best path is -550.7 nats, while the forward algorithm gives $\log P(x) = -533.4$ nats for the sum over all paths.

Several general lessons can be read from the figure:

- The Viterbi path is a "hard" answer that misses short loaded segments altogether. Switching into the loaded state costs $\log 0.05$, which a few sixes cannot pay for.
- The posterior is a "soft" answer. It rises in every loaded segment, even where Viterbi does not switch, and shows how uncertain the boundaries are.
- The probability of the single best path is much smaller than the total probability of the sequence: many paths contribute. When the question is "does this sequence come from this model?", the forward score is the right one to use, and this is what HMMER3 does.

The first rows of the dynamic-programming tables (natural logarithms) show the difference between taking the maximum (Viterbi, v) and the sum (forward, f) over predecessors. At the first roll both are identical; a six favours the loaded die. After that, the forward values are always at least as large as the Viterbi values.

t	roll	$\log v_F(t)$	$\log v_L(t)$	$\log f_F(t)$	$\log f_L(t)$	$P(L \mid x)$
1	6	-2.48	-1.39	-2.48	-1.39	0.67
2	3	-4.33	-3.79	-4.05	-3.78	0.58
3	6	-6.17	-4.59	-5.77	-4.53	0.55
4	5	-8.01	-7.00	-7.30	-6.93	0.39
5	3	-9.86	-9.41	-9.00	-9.30	0.31
6	2	-11.70	-11.82	-10.77	-11.63	0.28
7	3	-13.54	-14.22	-12.57	-13.92	0.28
8	4	-15.39	-16.63	-14.38	-16.13	0.32

1.6.5 Profile hidden Markov models

A **profile HMM** is an HMM whose architecture mirrors a multiple sequence alignment of a protein (or nucleic-acid) family (Krogh et al. 1994; Eddy 1998). Each consensus column of the alignment becomes a node with three states (Figure 1.9):

- a **match state** M_k, which emits a residue with column-specific probabilities $e_{M_k}(a)$. This is the HMM equivalent of one column of a PSSM;
- an **insert state** I_k, which emits residues between columns k and $k+1$, usually with background frequencies. Its self-transition $I_k \to I_k$ allows insertions of any length;
- a **delete state** D_k, a *silent* state that emits nothing and lets the sequence skip column k.

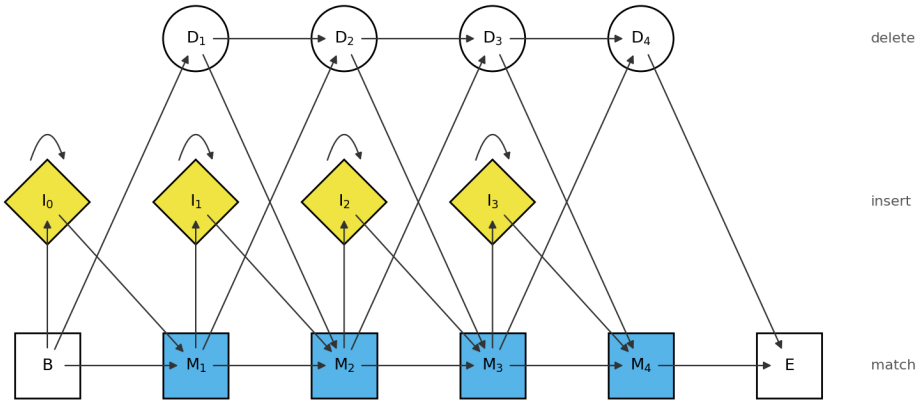


Figure 1.9: The core architecture of a profile HMM with four match states. Squares are match states (and the begin and end states), diamonds are insert states with self-loops, and circles are silent delete states. The transitions shown are those of HMMER's Plan 7 architecture, which has no transitions between insert and delete states.

A path through the model from begin to end is an alignment of the sequence to the family. Matches use M states, insertions relative to the consensus use I states, and deletions use D states. Unlike a PSSM with fixed gap penalties, the transition probabilities differ between positions. A loop region where insertions are common in the family gets a high $M_k \to I_k$ probability (a cheap gap), while a conserved

core gets very expensive gaps. With log-odds scoring this is equivalent to position-specific affine gap penalties: the gap-open cost is $\approx -\log a_{M_k I_k}$ (or $-\log a_{M_k D_{k+1}}$) and the extension cost $\approx -\log a_{I_k I_k}$ (or $-\log a_{D_k D_{k+1}}$).

> **Profiles, PSSMs, profile HMMs and pair HMMs are not the same**
>
> The terms are often used loosely, but they describe different models.
>
> - A **PWM/PSSM** gives a score for every residue at every position. A PWM for a motif has no gaps at all. The PSSM of PSI-BLAST is used with ordinary affine gap penalties that are the same at every position.
> - A **Gribskov profile** (and PROSITE's generalised profiles) adds position-specific gap penalties, but the scores need not come from a probabilistic model.
> - A **profile HMM** models a family with explicit match, insert and delete states. Residue emissions *and* insertion and deletion behaviour are position-specific probabilities, and scores (including the forward score summed over all alignments) have a probabilistic interpretation.
> - A **pair HMM** models the alignment of *two* sequences, with one match and two gap states (Section 1.3). It is the probabilistic version of Gotoh's algorithm and contains no family information.

1.6.5.1 Architecture: from Krogh's model to Plan 7

The model of Krogh et al. (1994) had nine transitions per node, including transitions between insert and delete states. HMMER2 and HMMER3 use the simpler **Plan 7** architecture, with seven transitions per node: $M \to M$, $M \to I$, $M \to D$, $I \to M$, $I \to I$, $D \to M$ and $D \to D$. Removing $I \leftrightarrow D$ transitions costs little, since an insertion directly followed by a deletion is rarely the best explanation of an alignment, and simplifies the algorithms.

Plan 7 also adds special states around the core model to control *how* a sequence may align to it:

- **N** and **C** states emit unaligned flanking sequence before and after the match to the model (with background composition);
- **B** and **E** are silent begin and end states of the core;
- a **J** ("joining") state allows the model to be used several times in one sequence. This is needed for proteins that contain several copies of the same domain.

Through the transitions out of B and into E, the same model can be used in different **alignment modes**. *Global* (with respect to the model) alignment requires the whole family model to be matched, from M_1 to M_K. *Local* alignment allows entry and exit at any match state, so that only a fragment of the model needs to be present. Local alignment with respect to both the model and the sequence corresponds to Smith–Waterman. Global with respect to the model but local with respect to the sequence is often called **glocal**. The N, C and J states make it *multi-hit* or *uni-hit*. HMMER3 searches use local (multi-hit) alignment by default, and the statistics for its forward scores are worked out for that mode.

1.6.5.2 Building a profile HMM from a multiple alignment

A profile HMM is usually estimated from a multiple sequence alignment (for example a Pfam seed alignment; Section 1.9) in a few steps. In HMMER this is done by `hmmbuild`.

1. **Choosing match columns.** Columns in which a sufficient fraction of sequences have a residue (by default at least half, after weighting) become match states. The remaining columns are treated as insertions.

2. **Counting.** With the columns assigned, the state path of every sequence is known, so emissions and transitions can be counted directly. No Baum–Welch is needed.
3. **Sequence weighting.** Alignments are biased: they often contain many near-identical sequences from well-studied organisms. Weighting schemes down-weight redundant sequences, for example position-based weights that give more weight to sequences with rare residues in a column. HMMER3 also reduces the *effective number of sequences* so that the model has a target information content. This stops a model built from thousands of sequences from becoming too specific to detect distant members.
4. **Priors.** Counts from a column with only a handful of sequences are poor estimates of the residue distribution of the whole family. They are combined with prior knowledge in the form of a **Dirichlet mixture prior**. Its components describe the typical kinds of alignment column (hydrophobic, aromatic, small, charged, conserved glycine, and so on). The posterior mean estimate of the emission probabilities is then $e_{M_k}(a) = (c_{ka} + \alpha_a) / \sum_{a'} (c_{ka'} + \alpha_{a'})$ for a single Dirichlet component with parameters α. With a mixture, each component is weighted by how well it explains the observed counts. This is a principled replacement for the pseudocounts of PSI-BLAST.

Profile HMMs can also be trained from *unaligned* sequences with Baum–Welch, which simultaneously produces a multiple alignment (Krogh et al. 1994). This was demonstrated on globins and kinases. In practice, alignment programs (Section 1.6) followed by HMM estimation work better, because Baum–Welch easily gets stuck in poor local optima.

1.6.5.3 Scoring and statistics

To decide whether a sequence belongs to a family, its probability under the profile HMM is compared with its probability under a **null model** R, usually a simple model that emits residues independently with background frequencies q_a (with a geometric length distribution). The score is the log-odds ratio, reported in **bits** (logarithm base 2):

$$S = \log_2 \frac{P(x \mid \text{HMM})}{P(x \mid R)}.$$

With the Viterbi path instead of the sum, this reduces to a sum of position-specific log-odds scores, $\log_2 \left(e_{M_k}(a)/q_a\right)$ for match emissions plus log transition probabilities, just like a PSSM with position-specific gap penalties. Because real sequences often have biased composition (low-complexity regions, transmembrane segments), HMMER also applies a second null model ("null2") that corrects scores for composition bias.

A bit score must be converted into a measure of significance (see Section 1.4 for the general theory). Viterbi scores of local alignments of random sequences follow a Gumbel (extreme-value) distribution, as for BLAST. Eddy showed that a probabilistic local alignment model has a fixed slope parameter $\lambda = \log 2$ when scores are in bits, both for Viterbi scores (Gumbel) and for forward scores (whose high-scoring tail is exponential) (Eddy 2008). Only the location parameter has to be estimated, which `hmmbuild` does with a short simulation. HMMER2 had needed a separate, slow calibration step (`hmmcalibrate`). This made it possible for HMMER3 to report E-values for **forward scores**. Summing over all alignments is more sensitive than using the single best one, because the alignment of a remote homologue is often uncertain. The E-value is, as in BLAST, the number of hits with at least this score expected by chance in a search of a database of this size.

1.6.5.4 HMMER and HMMER3

HMMER, written by Sean Eddy, is the standard implementation of profile HMMs. It is the engine behind Pfam, InterPro and many other family databases. HMMER1 appeared in the mid-1990s, and HMMER2 (1998) introduced Plan 7 and local alignment. Its main drawback was speed: a profile HMM search was 100–1000 times slower than BLAST, which limited it to searching the relatively small family databases. An independent profile HMM package, SAM (Sequence Alignment and Modeling system), was developed in parallel by the Santa Cruz group.

HMMER3 (Eddy 2011), a complete rewrite released in 2010, made profile HMM searches about as fast as BLAST without losing sensitivity. It uses a pipeline of increasingly expensive filters. Each target sequence must pass a score threshold, expressed as a P-value, to continue to the next stage:

1. **MSV filter** ("multiple segment Viterbi"). This is an ungapped, local, multi-hit alignment of the model to the sequence. It finds one or more high-scoring ungapped segments, similar in spirit to BLAST's ungapped extension. By default only about 2% of sequences ($P < 0.02$) pass.
2. **Viterbi filter**. This is a full gapped Viterbi alignment, computed with reduced numerical precision. About 0.1% of sequences ($P < 10^{-3}$) pass.
3. **Forward filter**. The full forward score in floating point, passed at $P < 10^{-5}$.
4. **Domain definition and alignment**. For the surviving sequences, forward–backward is used to compute posterior probabilities. From these HMMER identifies the individual domain hits and their boundaries, computes per-domain scores, and produces *maximum expected accuracy* alignments annotated with posterior probabilities.

The filters are fast because they are implemented with **SIMD vector instructions**. These process 16 one-byte scores (MSV) or 8 two-byte scores (Viterbi) in one instruction, using Farrar's "striped" layout of the dynamic-programming matrix, originally developed for Smith–Waterman (Farrar 2006). Because the early filters only need to be accurate enough to reject non-homologues, they can use reduced precision. The expensive full computation is done only for the small fraction of sequences that might be true hits.

The HMMER package contains programs for different search directions:

- `hmmbuild` builds a profile from an alignment;
- `hmmsearch` searches a profile against a sequence database;
- `hmmscan` searches a sequence against a profile database, for example annotating a protein with Pfam domains;
- `phmmer` searches a single sequence against a sequence database, using a profile built from the sequence and a substitution matrix;
- **`jackhmmer`** does iterative searches, like PSI-BLAST: each round builds a new profile HMM from the significant hits of the previous round. It was one of the programs used to build multiple alignments for AlphaFold2 (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021);
- `nhmmer` and `nhmmscan` search DNA and RNA;
- `hmmalign` aligns many sequences to a profile, which is how large alignments (such as Pfam full alignments) are built from a small seed.

1.6.5.5 Sequence and domain scores

HMMER output has two levels, which are often confused. The **per-sequence** score and E-value measure the evidence that a whole target sequence contains the family, summed over all its matches. The

per-domain scores measure each individual match. The two can disagree. A protein with several weak copies of a repeat can be significant as a sequence although no single copy is; a multidomain protein can have one strong domain and a weak, doubtful second hit.

Each domain hit has two E-values. The **independent E-value** treats the domain as if it were the only hit and uses the whole database as the search space. The **conditional E-value** asks how significant the domain is *given* that the sequence has already been accepted as a member, and uses only the small number of sequences that passed. The conditional E-value is therefore much smaller and is meant for deciding whether additional, weaker domains in an accepted sequence are real; it is not a measure of database-wide significance. A report should say which statistic and which threshold were used.

Both HMMER and PSI-BLAST also separate a **reporting threshold** (which hits are shown) from an **inclusion threshold** (which hits count as members, and in iterative searches are used to build the next model). Changing the first only changes the output; changing the second changes the next profile, which is how profile drift begins.

1.6.5.6 Profile–profile comparison: the HH-suite

A profile HMM is more sensitive than a single sequence because it knows which positions are conserved and where gaps are likely. The logical next step is to represent *both* the query and the database entries as profile HMMs and align the two models. Johannes Söding's **HHsearch** (Söding 2004) does this. It aligns two profile HMMs by a Viterbi-like dynamic programming over pairs of states. The score for aligning column i of the query with column j of the template is the log of the probability that the two match states emit the same residue (their *co-emission* probability), relative to the background:

$$S_{aa}(q_i, t_j) = \log \sum_a \frac{q_i(a)\, t_j(a)}{f(a)},$$

where $q_i(a)$ and $t_j(a)$ are the emission probabilities and $f(a)$ the background frequencies. Optionally, a score for agreement between predicted and observed secondary structure is added. HHsearch, and the HHpred server built on it (Soding et al. 2005), became the most sensitive methods for detecting remote homologues of known structure, and HHpred-based servers performed among the best template-based modelling methods in several CASP experiments.

HHblits (Remmert et al. 2011) made the approach iterative and fast enough to search the whole sequence space. The database is first clustered (for example UniProt into clusters at about 20–30% identity: uniprot20, later Uniclust30 (Mirdita et al. 2016), and the metagenomic BFD). Each cluster is represented by a profile HMM built from its alignment. HHblits then searches this HMM database with a query HMM in a PSI-BLAST-like iterative way. Each round adds the sequences of the significant clusters to the query alignment. A two-stage **prefilter**, with fast ungapped and then gapped comparisons of discretised profile columns, reduces the tens of millions of database HMMs to a few thousand before the expensive HMM–HMM alignment. HHblits builds multiple alignments that are both deeper and more accurate than those from PSI-BLAST, in a fraction of the time. **HH-suite3** (Steinegger et al. 2019) vectorised the HMM–HMM Viterbi and forward algorithms with SIMD instructions, much as HMMER3 did for sequence–profile comparison.

HH-suite tools are central to protein structure prediction (Chapter 9):

- HHsearch against a database of PDB-derived profiles (PDB70) finds templates for homology modelling;

- HHblits against Uniclust30 and BFD was used, together with jackhmmer, to construct the multiple sequence alignments that are the main input of AlphaFold2 (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021).

The deep, accurate alignments needed for coevolution analysis and deep learning are therefore a direct product of profile HMM methods. Today, much of this work is done with faster MMseqs2-based pipelines (Section 1.4).

1.6.5.7 How much do profiles gain?

Benchmarks on structurally classified proteins, where homology can be established from structure even when sequence similarity is undetectable, give a consistent picture. In a classic comparison, Park et al. (1998) showed that methods using multiple sequences (PSI-BLAST and the HMM program SAM-T98) detected about three times as many remote homologues as pairwise methods at the same error rate. Profile–profile methods add another substantial gain over sequence–profile methods (Söding 2004). The resulting ordering is

$$\text{sequence–sequence} \;<\; \text{sequence–profile} \;<\; \text{profile–profile}.$$

The gain comes from more information on both sides, not from the HMM formalism alone. A PSSM built from a good alignment can come close to a profile HMM. The HMM's advantages are position-specific gap probabilities, principled priors, the forward score and calibrated statistics.

1.6.5.8 Limitations

- **Independence of positions.** A profile HMM assumes that columns evolve independently given the path. Correlations between positions, such as compensating mutations of residues in contact, are invisible to it. Models that capture them, such as Potts models and direct-coupling analysis, are the basis of coevolution-based contact prediction (Chapter 9). Protein language models (Section 1.10) learn such dependencies implicitly.
- **Simple length distributions.** Self-loops give geometric distributions of insertion lengths, and a first-order Markov chain has no memory beyond the previous state. Gene finders use *generalised* HMMs with explicit length distributions for exons for this reason (Chapter 5).
- **Dependence on the alignment.** A profile HMM is only as good as the multiple alignment it is built from. Misaligned regions or contaminating non-homologous sequences degrade it.
- **Profile drift.** In iterative searches (jackhmmer, HHblits, PSI-BLAST), a single false positive included in one round can pull the profile towards an unrelated family. The search then reports many confident but wrong hits. Inclusion thresholds should therefore be conservative, and hits near the threshold should be checked (see the warning signs listed under PSI-BLAST above).
- **Local optima in training.** Baum–Welch finds local maxima only, which is why profiles are usually estimated from fixed alignments.

Questions to think about

- Why is the forward score a better measure of homology than the Viterbi score for remote homologues, and why did it take until HMMER3 to use it in database searches?
- How do the transition probabilities of a profile HMM relate to the gap penalties of Smith–Waterman? What can a profile HMM express that affine gap penalties cannot?

- In what sense is a PSSM a special case of a profile HMM?
- Why does an iterative search (PSI-BLAST, jackhmmer, HHblits) find more homologues than a single search, and what can go wrong?
- Why is profile–profile comparison more sensitive than sequence–profile comparison?

Sources

Parts of this section are adapted from the Wikipedia articles Hidden Markov model, Viterbi algorithm, Forward algorithm, Forward–backward algorithm, Baum–Welch algorithm, HMMER and HH-suite (CC BY-SA 4.0, retrieved 2 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Sequence Databases, BLAST, and Profile HMMs" (CC BY 4.0), used with permission.

1.7 RNA homology search

The methods discussed so far work well for proteins, and for DNA that is very similar. Many genes, however, do not code for proteins. Their products are *functional* non-coding RNAs (ncRNAs). Examples are transfer and ribosomal RNAs, the RNA components of RNase P and the signal recognition particle, small nuclear and small nucleolar RNAs, microRNA precursors, riboswitches and many regulatory RNAs in bacteria and viruses. These molecules are among the oldest in the cell: ribosomal RNA is conserved across all of life and was the basis of the three-domain tree (Chapter 4). Yet ncRNA genes are notoriously hard to find by sequence similarity. This section explains why, and how the problem was largely overcome by extending profile HMMs (Section 1.6) to models that also describe RNA secondary structure.

1.7.1 Why RNA is different

The function of most ncRNAs depends on their *structure*, and the structure is formed largely by base pairing within the molecule. Complementary segments pair to form double-helical **stems**, connected by single-stranded **hairpin loops**, **internal loops**, **bulges** and **multibranch loops**. Besides the Watson–Crick pairs G–C and A–U, the **G–U wobble pair** is common. The set of base pairs is the molecule's **secondary structure**. It is usually written in *dot-bracket* notation, where matching brackets mark paired bases and dots mark unpaired ones, for example `((((...))))` for a simple hairpin.

Three properties make RNA homology search harder than protein search:

- **In structured RNAs, structure is conserved better than sequence.** In RNAs whose function depends on a conserved fold (tRNAs, rRNAs, RNase P RNA, riboswitches and many other structured ncRNAs), selection acts largely on the base pairs, not on the individual bases. A G–C pair can change to A–U, C–G or G–U with little effect on the structure. Such **compensatory** (or *covarying*) mutations change two positions of the sequence at once, while the structure stays the same. Two homologous RNAs can therefore share the same fold with very low sequence identity. A sequence-only method sees two mismatches where the evolutionary signal is in fact a conserved base pair. This is not a property of all RNA genes. Where function depends on sequence motifs, for example the seed region of a microRNA or a protein-binding site, sequence can be the more conserved feature, and some ncRNAs conserve neither (see *Limitations* below).
- **A small alphabet.** With four nucleotides, a random position matches about a quarter of the time (1). Short ncRNAs of 70–300 nucleotides therefore give little statistical signal in a pairwise comparison.

- **Long-range dependencies.** The two halves of a stem can be hundreds of nucleotides apart in the sequence. A profile HMM treats positions as independent given the state path, so it cannot represent the fact that position i must be complementary to position j.

For close homologues, sequence-based tools work well enough: BLASTN, or profile HMMs for DNA such as nhmmer (Wheeler and Eddy 2013). For distant homologues, the structure has to be part of the model.

1.7.2 Predicting RNA secondary structure

Structure-aware search builds on algorithms for RNA folding, which are themselves dynamic programming (Section 1.3). The key simplification is that most base pairs in a secondary structure are **nested**: if i pairs with j and k pairs with l, with $i < k$, then either k and l both lie between i and j, or both lie outside. Nested structures can be decomposed recursively, because the pair (i, j) splits the sequence into an inside part and an outside part that do not interact. Structures that violate this rule are called **pseudoknots**. They occur in real RNAs, but they are usually left out of the models because they make the algorithms much more expensive.

- **Nussinov's algorithm** (Nussinov and Jacobson 1980) finds the structure with the maximum number of base pairs. The best structure of the subsequence $x_i \dots x_j$ is computed from the best structures of shorter subsequences: i is unpaired, j is unpaired, i pairs with j, or the subsequence splits into two independent parts at some k. Filling the table takes $O(L^3)$ time and $O(L^2)$ memory for a sequence of length L.
- **Minimum free energy (MFE) folding** (Zuker and Stiegler 1981) replaces "count the pairs" with a thermodynamic model. Most of the stability of a stem comes from *stacking* of adjacent base pairs, and loops cost energy depending on their type and size. The parameters (*nearest-neighbour* energies) are measured experimentally. The recursion has the same structure and the same $O(L^3)$ complexity. This is the approach of programs such as mfold and the ViennaRNA package (`RNAfold`). A refinement computes the partition function, and from it the probability of every possible base pair, instead of a single optimal structure.

Single-sequence prediction is far from perfect. Many structures have similar energies, the energy parameters are approximate, and in the cell proteins and other RNAs influence the fold. Only a minority of predicted structures are correct in every pair, especially for long RNAs.

1.7.3 Comparative analysis: covariation

The most reliable evidence for an RNA structure comes from evolution. If an alignment of homologous RNAs shows that positions i and j always form a Watson–Crick or G–U pair, even though both positions vary, they are almost certainly paired. This **comparative** approach was used to determine the secondary structures of ribosomal RNAs and other large RNAs long before their crystal structures were known, and it proved remarkably accurate. Covariation can be measured with *mutual information* between alignment columns, or combined with folding energies to predict a consensus structure for the whole alignment, as in RNAalifold (Hofacker et al. 2002). It is the RNA counterpart of the coevolution analysis used for protein contact prediction (Chapter 9).

Covariation is only evidence if it exceeds what is expected from the phylogeny of the sequences alone. Closely related sequences share many substitutions simply by descent. **R-scape** (Rivas et al. 2016) tests whether the covariation of each proposed pair is statistically significant, given the phylogeny and the amount of variation in the alignment. Applied to published structures, it confirmed the pairs of well-established RNAs such as tRNA and rRNA. However, it found no significant support for the

conserved structures proposed for several long non-coding RNAs. This is a useful reminder that a predicted structure, even in an alignment, is a hypothesis.

A harder problem is that structural alignment and structure prediction depend on each other. A correct structure requires a correct alignment, and for divergent RNAs a correct alignment requires knowing the structure. Sankoff (1985) gave a dynamic-programming algorithm that solves both simultaneously, folding and aligning two sequences at once. Its cost, $O(L^6)$ time and $O(L^4)$ memory for two sequences, makes it impractical in its full form. Tools such as LocARNA (Will et al. 2007) use restricted versions that work with base-pair probabilities and are fast enough to cluster thousands of ncRNA candidates into families.

1.7.4 From HMMs to stochastic context-free grammars

The step from profile HMMs to structure-aware models is easiest to see in the language of formal grammars. A **grammar** is a set of production rules that generate strings. In the Chomsky hierarchy, the simplest class, *regular grammars*, can only generate a string from left to right, with each step depending on the current state. HMMs are probabilistic regular grammars. **Context-free grammars** allow rules that emit symbols at both ends of a substring at once. For example,

$$S \to \mathtt{a}\,S\,\mathtt{u} \mid \mathtt{g}\,S\,\mathtt{c} \mid \ldots \mid \varepsilon$$

generates nested complementary pairs, which is exactly the structure of an RNA stem. A **stochastic context-free grammar** (SCFG) attaches a probability to every rule. A *parse* of a sequence, the series of rule applications that generates it, then corresponds to a secondary structure, and its probability is the product of the rule probabilities.

The algorithms for SCFGs mirror those for HMMs (Section 1.6):

HMM	SCFG	Purpose
Viterbi	CYK (Cocke–Younger–Kasami)	most probable parse, i.e. structure or structural alignment
Forward	Inside	total probability of the sequence
Forward–backward	Inside–outside	posterior probabilities; training by EM

The price of modelling pairs is complexity. Where an HMM fills a table indexed by sequence position, an SCFG fills a table indexed by *subsequences* (i, j). For a fixed grammar, the time is $O(L^3)$ instead of $O(L)$.

1.7.5 Covariance models

A **covariance model** (CM) is to an RNA family what a profile HMM is to a protein family. It is a profile SCFG built from a multiple alignment that is annotated with the family's consensus secondary structure (Eddy and Durbin 1994). The consensus structure is converted into a tree of **nodes**. There are nodes for base pairs, for unpaired positions on the left or right, and *bifurcation* nodes where the structure branches into two stems, as in a multibranch loop. Each node is expanded into states, much like the match, insert and delete states of a profile HMM:

- a **pair state** emits two nucleotides at once, with 16 probabilities for the 16 possible pairs. This is where covariation is captured: a family in which a position pair is G–C in some members and A–U in others gives high probability to both pairs, and low probability to mismatches such as G–A, even though each position on its own is variable;
- states for single-stranded positions emit one nucleotide;
- insert and delete states allow insertions and deletions relative to the consensus, with position-specific probabilities.

Parameters are estimated from the alignment with sequence weighting and Dirichlet priors, as for profile HMMs (Section 1.6). Sequences are scored by log-odds against a null model. Searching a genome means aligning the model to every window of the sequence. This scales roughly as the cube of the window length for every position of the genome, which made the first CM searches extremely slow.

1.7.6 Infernal

Infernal ("INFERence of RNA ALignment"), developed by Eric Nawrocki and Sean Eddy, is the standard implementation of covariance models. It plays the same role for RNA as HMMER does for proteins. The main problem was speed, and it was addressed in steps:

- **Banded dynamic programming.** *Query-dependent banding* precomputes, for each state of the model, the range of subsequence lengths it can plausibly generate, and restricts the CYK and inside computations to these bands (Nawrocki et al. 2009). Bands can also be derived from a profile HMM alignment of the same sequence.
- **HMM filters.** Infernal 1.1 (Nawrocki and Eddy 2013) adopted the filter pipeline of HMMER3 (Section 1.6). A profile HMM built from the same family first removes almost all of the genome using the fast MSV, Viterbi and forward filters. The full CM, with its much higher sensitivity, is only run on the remaining candidate regions. This made searches about 100 times faster than in Infernal 1.0.
- **Statistics.** As in HMMER, scores are reported in bits with E-values calibrated on random sequences. Local alignment modes allow hits that cover only part of the model, including *truncated* hits at the ends of sequencing reads or contigs.

With these improvements, searching a whole bacterial genome for all known RNA families, or a CM against large sequence databases, became routine.

1.7.7 Rfam

Rfam is the database of RNA families, designed explicitly as the RNA counterpart of Pfam (Section 1.9). It was started in 2003 at the Sanger Institute, by Sam Griffiths-Jones, Alex Bateman, Sean Eddy and colleagues (Griffiths-Jones et al. 2003), and is now maintained at EMBL-EBI (Kalvari et al. 2020; Ontiveros-Palacios et al. 2024). Each family has:

1. a curated **seed alignment** of representative sequences, annotated with the consensus secondary structure (in Stockholm format);
2. a **covariance model** built from the seed with Infernal;
3. a **full alignment** of all sequences in the database that score above the family's manually chosen gathering threshold.

The same design as Pfam means that only the seed has to be curated, while the full alignments are regenerated automatically. Related families are grouped into clans. Like Pfam, Rfam links its family descriptions to Wikipedia, and many Wikipedia articles on RNA families were written in this collaboration. The database has grown from 25 families in the first release to several thousand today. Rfam 13

reorganised the database around complete reference genomes, and recent releases have added many families from viruses, metagenomes and microRNAs. Until release 12, the slowness of CMs meant that candidates were first found by BLAST, but Infernal 1.1 made this filter unnecessary.

Rfam is used to annotate ncRNA genes in newly sequenced genomes, much as Pfam and InterPro are used for proteins. Its main difficulty is the opposite of Pfam's. Eukaryotic genomes contain many *pseudogenes* and repeats derived from ncRNAs, for example from tRNAs and small nuclear RNAs, which match the models well but are not functional genes. Separating these from the real genes requires additional evidence. Introns, which occur in some ncRNA genes, are also not described by the models.

1.7.8 Specialised tools: tRNAs

For the best-studied RNA families, specialised programs combine CMs with extra knowledge. **tRNAscan-SE** (Lowe and Eddy 1997; Chan et al. 2021) finds tRNA genes in genomes by first using fast filters, and then scoring the candidates with covariance models of the tRNA cloverleaf. The cloverleaf, with its acceptor stem, D arm, anticodon arm and T arm, is the classic example of a structure that is much better conserved than its sequence. tRNAscan-SE also identifies the anticodon and hence the amino acid specificity, and flags probable pseudogenes. It is used in essentially every genome annotation pipeline (Chapter 5).

1.7.9 Limitations

- **Pseudoknots and tertiary interactions** cannot be represented by context-free models, so CMs ignore them. This loses some information, for example in riboswitches and RNase P RNA, where pseudoknots are conserved.
- **Cost.** Even with filters, CM searches are much slower than profile HMM searches.
- **Dependence on the seed alignment.** A CM is only as good as the structural alignment and consensus structure it is built from.
- **Not all ncRNAs are structured.** Many long non-coding RNAs show little conservation of sequence or structure, and their function may depend on short sequence motifs or on the act of transcription itself. For them, structure-based homology search has little to offer, and homology can often only be inferred from conserved genomic position (synteny).

See also: further explanation in Section D.1.6; related Wikipedia articles in Section E.1.6.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Rfam, Probabilistic context-free grammar, Nucleic acid structure prediction and Non-coding RNA (CC BY-SA 4.0, retrieved 2 October 2026).

1.8 Whole-genome alignment and conservation

The methods earlier in this chapter align genes and proteins: sequences of a few hundred to a few thousand residues, whose homology is usually one-to-one along their length. Aligning whole genomes is a different problem, and its results, multiple genome alignments and the conservation scores computed from them, are among the most used tracks in genome browsers (Section 2.4).

1.8.1 Why genomes are not long genes

A genome alignment should identify, for every base, the homologous base (if any) in each of the other genomes. Several things make this harder than aligning two genes:

- **Rearrangements.** Inversions, translocations, fusions and fissions break the collinearity of genomes (Section 6.5). A single global alignment, as computed by Needleman–Wunsch, cannot represent a block that has moved or changed orientation.
- **Duplications and losses.** A region may have one copy in one genome and several in another, through segmental or whole-genome duplication. The alignment must then decide which copy is the orthologue, or allow one base to align to several.
- **Repeats.** Transposable elements and other repeats make up about half of mammalian genomes. They create huge numbers of short, spurious matches, so repeats are usually soft-masked (written in lower case) and used only to extend alignments, not to start them.
- **Size and divergence.** Mammalian genomes are billions of bases long, and most of their sequence is neutrally evolving, so between distant species only a small part can be aligned with confidence.

Whole-genome alignment is therefore done in stages: find local matches, chain them into collinear blocks, and decide which blocks represent orthology.

1.8.2 Pairwise alignment: BLASTZ/LASTZ, chains and nets

BLASTZ was developed to align the human and mouse genomes (Schwartz et al. 2002). It follows the seed-and-extend scheme of BLAST (Section 1.4), but with *spaced seeds* (by default 12 matching positions in a window of 19, with one transition allowed), which are more sensitive for diverged non-coding DNA than contiguous words. The best seed is not fixed: it depends on the length and similarity of the matches that are sought, so seed patterns and word sizes are search parameters to be tuned rather than constants (Gotea 2003). The seeds are followed by gapped extension and a second, more sensitive search between the anchors that were found. Its successor **LASTZ** is used in the same way.

The local alignments are then organised in two steps, introduced at UCSC (Kent et al. 2003):

- A **chain** is an ordered series of gapless blocks that are in the same order and orientation in both genomes. Gaps are allowed in *both* sequences between blocks, which is how insertions, deletions and unalignable sequence appear. Chains are found by dynamic programming over the local alignments, much like the chaining of seeds in minimap2.
- A **net** is a hierarchical selection of chains in which each base of the reference genome is covered at most once. The best chain is placed first, and gaps in it are filled with the best chains that fit there. Nets therefore distinguish the orthologous, collinear alignment from alignments to paralogous copies, and they show inversions and other rearrangements as nested chains.

Chains are also the basis of **liftOver**, the tool that converts coordinates between assemblies of the same genome (Section 2.4).

1.8.3 Multiple genome alignment

Pairwise alignments to one genome can be combined into a multiple alignment. **MULTIZ**, part of the threaded blockset aligner (Blanchette et al. 2004), progressively merges pairwise alignments that all share a reference genome, following a phylogenetic tree. The UCSC multiple alignments of vertebrates and mammals are built this way. Such alignments are *reference-based*: they only contain sequence that aligns to the reference (for example human), and sequence absent from the reference is lost.

Reference-free methods treat all genomes symmetrically. In Ensembl, **Enredo** builds a graph of collinear segments from anchors in all genomes, allowing for duplications and rearrangements, and **Pecan** aligns each segment with a consistency-based method in the spirit of T-Coffee (Section 1.6) (Paten et al. 2008). **Progressive Cactus** (Armstrong et al. 2020) represents the alignment as a *cactus graph*, in which rearrangements and duplications are natural, and aligns genomes progressively along a guide tree, reconstructing an ancestral genome at each internal node. Because each new genome only has to be aligned to its nearest ancestors, the method scales to hundreds of genomes. It was used to build the Zoonomia alignment of about 240 mammal species (Genereux et al. 2020; Christmas et al. 2023) and alignments of hundreds of bird genomes.

1.8.4 Evaluating genome alignments

There is no ground truth for real genomes, so genome aligners are hard to evaluate. The **Alignathon** competition (Earl et al. 2014) compared aligners on genomes simulated with a known evolutionary history, including rearrangements and duplications, where precision and recall of aligned base pairs can be computed, and on real fly genomes, where only indirect measures are possible. Methods differed substantially, especially for distant genomes, and no method was best on all measures. The lesson matches that for protein alignment (Section 1.11): a genome alignment is an estimate, and downstream results, such as which bases are conserved, inherit its errors.

1.8.5 Conservation scores

Given a multiple genome alignment and a phylogenetic tree with a *neutral model* of substitution, usually estimated from sites assumed to evolve without constraint, such as fourfold degenerate sites or ancient repeats, each column can be asked: did it change less than expected? Purifying selection removes mutations at functional positions, so a deficit of substitutions is evidence of function, whether or not the sequence codes for protein. Three scores are widely used:

- **phastCons** (Siepel et al. 2005) uses a two-state **phylo-HMM** (Section 1.6): a hidden Markov model along the genome whose states emit alignment columns according to a phylogenetic model. In the *conserved* state all branch lengths are scaled down by a factor $\rho < 1$. The posterior probability of the conserved state is reported for each base, and runs of high probability are called *conserved elements*. Because neighbouring bases share a state, phastCons finds elements rather than single positions.
- **phyloP** (Pollard et al. 2009) tests each base (or element) separately for a departure from the neutral rate, and reports $-\log_{10} p$-values: positive for conservation (slower than neutral) and negative for *acceleration* (faster than neutral). It can also test one lineage, for example the human branch, which is how **human accelerated regions**, conserved in other mammals but rapidly changed in humans, were found (Pollard et al. 2006).
- **GERP** (Cooper et al. 2005) measures constraint as **rejected substitutions**: the number of substitutions expected at a column under neutrality minus the number estimated to have occurred. A score of 4 means that four substitutions that would have happened have been removed by selection. GERP++ (Davydov et al. 2010) calls constrained elements from these scores.

Conservation scores are used to find functional non-coding elements, such as enhancers and non-coding RNAs, and to estimate how much of the genome is under constraint (Section 6.9). They are also among the most informative features for interpreting variants: a variant at a strongly constrained base is more likely to be harmful, and base-level constraint from the Zoonomia alignment helps to prioritise non-coding variants in rare disease and association studies (Sullivan et al. 2023). Scores such as CADD include phyloP and GERP among their features (Section 10.6), and variant calling itself is described in Section 5.4.

Conservation scores also have important limitations:

- **They depend on the alignment.** Misaligned paralogues, assembly errors and missing data in some species look like substitutions or conservation. Repeats are often poorly aligned and therefore under-scored.
- **They need enough evolution.** Single-base resolution requires a large total branch length in the tree. A handful of primates cannot show that a single base is constrained; hundreds of mammals can. Alignments of distant species, on the other hand, contain only the most conserved elements.
- **Function can be lineage-specific.** Elements that are new or turn over quickly, such as many transcription factor binding sites, are not conserved across the tree although they are functional in one lineage. Lack of conservation is therefore not evidence of lack of function.
- **Not all deviation is selection.** Mutation-rate variation and GC-biased gene conversion can make a region look conserved or accelerated, as discussed for some human accelerated regions.

A conserved base is evidence that *some* function has been maintained by selection over the depth of the tree used. It does not say what the function is, or whether it is the same in humans.

See also: further explanation in Section D.1.10; related Wikipedia articles in Section E.1.10.

> **Sources**
>
> Written for this book from the primary literature cited in the text.

1.9 Databases and domains

So far we have mostly treated a protein as one sequence to be aligned from end to end, or at least as one region of local similarity. Real proteins are often built from several independent parts, *domains*, that have their own evolutionary histories. A protein of 800 residues may share one domain with a kinase, another with a transcription factor, and a third with nothing in the databases at all. Understanding domains is therefore essential for interpreting search results. It also explains why the most widely used annotation resources, Pfam and InterPro, are organised around domains and families rather than whole proteins. This section covers what a domain is, how domain families are built and searched (Pfam, Pfam-N and InterPro), and gives a first look at the structural classifications (SCOP, CATH, ECOD and TED) that are discussed further in Chapter 2 and Chapter 9.

1.9.1 What is a protein domain?

1.9.1.1 History of the concept

The idea of the domain grew out of the first protein crystal structures. In 1973 Donald Wetlaufer surveyed the structures then available, including hen lysozyme, papain and fragments of immunoglobulins produced by limited proteolysis. He noted that many proteins consist of separate, compact globular regions that are connected only by a few strands of polypeptide chain (Wetlaufer 1973). He proposed that such regions are *stable units of protein structure that can fold autonomously*, and suggested that folding of a large protein proceeds by nucleation within these regions. The immunoglobulins were a particularly clear case. Proteolysis cut the antibody into fragments that kept their structure and binding activity, and the sequence of each chain could be seen to consist of repeats of about 110 residues, each forming one compact “immunoglobulin domain”.

As more structures were determined, the term came to be used in three overlapping senses:

- **Structural domain:** a compact, globular sub-structure with more contacts within itself than with the rest of the protein, and usually with its own hydrophobic core.
- **Folding domain:** a unit that can fold, and remain stable, on its own when expressed or cut out of its parent protein.
- **Evolutionary (and functional) domain:** a unit that recurs in different proteins and different combinations. It has been moved around as one piece by gene duplication, fusion and recombination during evolution, and often carries a recognisable function, such as binding a ligand or catalysing a reaction.

In most cases these definitions agree: a compact structural unit found in many different proteins usually also folds on its own. But they need not agree, and different databases make different choices. This is one reason why domain boundaries and domain counts differ between resources.

1.9.1.2 Size and structure of domains

Domains are typically 50–250 residues long. Individual structural domains range from about 36 residues (in E-selectin) to almost 700 (in lipoxygenase), but about 90% are shorter than 200 residues, and the average is around 100 residues. Very small domains, below about 40 residues, are usually stabilised by metal ions or disulphide bridges; zinc fingers are the classic example. Domains longer than about 300 residues often turn out to contain more than one hydrophobic core and can be split further.

Each domain is built from secondary structure elements packed around a core. Domains are traditionally grouped into four classes by their secondary-structure content: *all-α*, *all-β*, *α/β* (mostly parallel β-sheets with β-α-β units, as in the TIM barrel or the Rossmann fold) and *α+β* (helices and mostly antiparallel strands in separate regions). These classes form the top level of both SCOP and CATH (below).

A domain need not be a single continuous stretch of sequence. When one domain has been inserted into a loop of another during evolution, the host domain is made up of two or more segments of the chain. Pyruvate kinase is the textbook example: its TIM-barrel domain is interrupted by an inserted β-barrel domain. About a quarter of structural domains are discontinuous in this way. A practical consequence is that a domain cannot always be represented by a single start and end position in the sequence, which complicates both sequence-based domain databases and automatic domain assignment.

Long-range interactions also matter in the other direction. Some proteins form oligomers by *domain swapping*, where a secondary-structure element or a whole domain of one chain replaces the equivalent part of a neighbouring chain. And in many multidomain proteins the domains are connected by flexible, often intrinsically disordered linkers, so their relative orientation varies. This is one reason why structure-based domain assignment, and structure prediction (Chapter 9), is usually more reliable for individual domains than for whole chains.

1.9.2 Multidomain proteins and domain architectures

Multidomain proteins are common, especially in eukaryotes. Estimates depend on the definitions and databases used, but roughly 40% of prokaryotic proteins and 65–80% of eukaryotic proteins are multidomain proteins, and the fraction is highest in multicellular animals (Apic et al. 2001). The linear order of domains along the chain, from N- to C-terminus, is called the protein's **domain architecture**.

Figure 1.10 shows three classic examples:

- **Src tyrosine kinase** has an N-terminal unique region followed by an SH3 domain, which binds proline-rich peptides; an SH2 domain, which binds phosphotyrosine; and a C-terminal kinase domain. The same SH2 and SH3 domains occur, in other combinations, in hundreds of other signalling proteins.

- **Tissue plasminogen activator** consists of a fibronectin type-I ("finger") module, an EGF-like module, two kringle domains and a trypsin-like serine protease domain. The extracellular proteins of blood clotting, fibrinolysis and complement are built from a small set of such modules, combined in many different ways. They were among the first examples of *domain shuffling*, the reuse of domains through exon shuffling and recombination.
- **The purine-biosynthesis enzymes** GARs, AIRs and GARt are three separate proteins in bacteria. In yeast, GARs and AIRs are fused into one chain while GARt is separate, and in vertebrates all three are fused into a single trifunctional polypeptide. The same activities are present in all three, but the way they are packaged into chains differs.

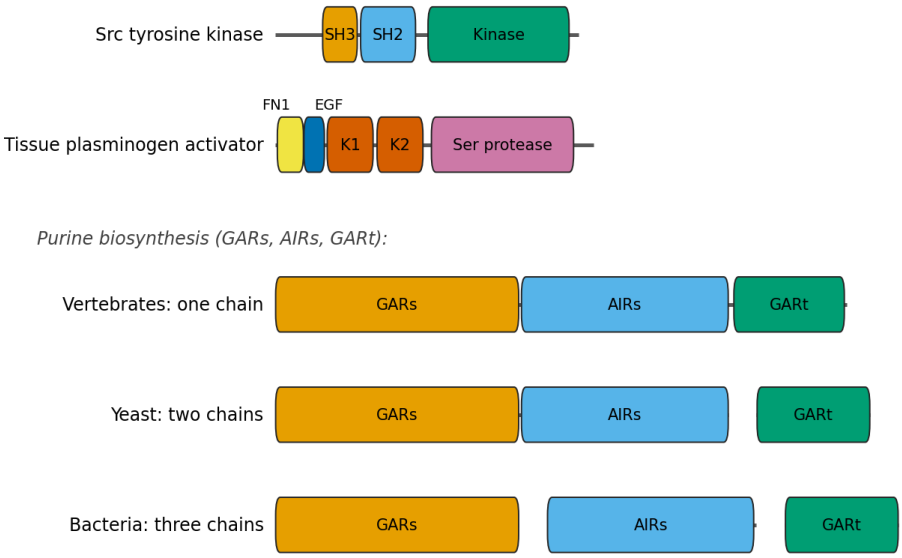


Figure 1.10: Schematic domain architectures (approximately to scale). Top: Src kinase and tissue plasminogen activator, two multidomain proteins built from widely reused domains (K1, K2: kringle domains; FN1: fibronectin type-I module). Bottom: the purine-biosynthesis enzymes GARs, AIRs and GARt are separate chains in bacteria, partly fused in yeast and fully fused in vertebrates.

The last example illustrates a general pattern. Many domains that form parts of multidomain proteins in eukaryotes exist as separate proteins in prokaryotes, which suggests that multidomain proteins have often arisen by *gene fusion* of previously independent genes. The opposite process, *gene fission*, also occurs. Because fused proteins usually function together, the observation that two proteins are fused in another genome is used to predict functional links ("Rosetta stone" proteins, Chapter 10).

1.9.2.1 How domain architectures evolve

Domain architectures change through a limited set of genetic mechanisms: duplication of a whole gene or part of it, fusion and fission, insertion of one domain into another, exon shuffling (in eukaryotes, where introns make recombination between domains easier), and transposition, including horizontal transfer between species. Several broad patterns have been found by comparing architectures across genomes:

- **Combinations are limited.** Although thousands of domain families are known, only a small fraction of all possible pairs of domains is ever observed together, and most domains combine

with only one or two partners. A few *promiscuous* domains, such as SH3, SH2, PH and the immunoglobulin domain, combine with very many others (Basu et al. 2008). Domain combinations also tend to keep their N-to-C order (Apic et al. 2001).

- **Most changes happen at the ends.** When architectures of related proteins are compared, most differences are the gain or loss of single domains at the N- or C-terminus, rather than insertions in the middle (Björklund et al. 2005).
- **Tandem repeats are common.** The simplest architecture is one domain repeated many times. Titin, a muscle protein of about 30,000 residues, contains more than a hundred immunoglobulin and fibronectin type-III domains. In some repeats, as in the two β-barrels of the trypsin-like serine proteases, the copies have diverged so far that their common origin is only visible in the structure.
- **Assembly mirrors evolution.** For protein complexes, the order in which subunits assemble often reflects the order in which they were added during evolution (Levy et al. 2008).

These patterns, and how domain rearrangements relate to gene duplication and to the evolution of new functions, are discussed further in Chapter 3.

1.9.2.2 Why domains matter for sequence searching

Domains have direct consequences for the search methods in this chapter:

- **Hits are local.** A significant hit between two multidomain proteins usually means that *one region* is homologous, not the whole chain. Local alignment (Section 1.3) will report the shared domain, but the rest of the two proteins may be unrelated.
- **Annotation transfer can go wrong.** If the query shares only an SH2 domain with a kinase, transferring the annotation "tyrosine kinase" is a mistake. Many errors in sequence databases have arisen in exactly this way and have then propagated as other proteins were annotated from the erroneous entries (Chapter 10).
- **Homology is not transitive across domains.** If protein A matches B through domain 1, and B matches C through domain 2, then A and C are not shown to be related at all. Iterative methods such as PSI-BLAST, and sequence clustering by single linkage, can "walk" from one family to another through multidomain proteins in this way. This is a common source of profile corruption (Section 1.6).
- **Search per domain.** For these reasons, sequences are often annotated by searching them against libraries of domain models, such as Pfam, instead of (or as well as) against whole sequences. Boundaries found in that way can then be used to cut a protein into domains before further analysis.
- **Low-complexity and repeat regions** (compositionally biased segments, coiled coils, transmembrane helices, short repeats) can produce high-scoring but non-homologous matches. They are often masked before searching, and are handled specially in domain databases.

1.9.3 Pfam

Pfam is the most widely used database of protein domains and families. It contains a curated multiple sequence alignment and a profile hidden Markov model (Section 1.6) for each family, and it is used to annotate the protein content of essentially every newly sequenced genome.

1.9.3.1 History

Pfam was started in 1995 by Erik Sonnhammer, Sean Eddy and Richard Durbin at the Sanger Centre in Cambridge, and was first described in 1997 (Sonnhammer et al. 1997). A major motivation was the

annotation of the first animal genome, that of the nematode *Caenorhabditis elegans*. The project was also inspired by Cyrus Chothia's argument (Chothia 1992) that most proteins belong to a limited number of families, perhaps around a thousand. If so, a manageable collection of family models could annotate a large fraction of all proteins. The estimate turned out to be too low: Pfam release 37.0 (June 2024) contains almost 22,000 families. But the main idea, that a library of family models gives broad and sensitive coverage, has been borne out. Already around release 29, about three quarters of all sequences in UniProtKB had at least one Pfam match.

Pfam was originally a consortium with sites in the UK, the USA and Sweden (Erik Sonnhammer, who now works at Stockholm University and SciLifeLab). Between 2012 and 2014 it moved to the European Bioinformatics Institute (EMBL-EBI), and since about 2022 it has been part of InterPro, through whose website it is now accessed (Mistry et al. 2020; Paysan-Lafosse et al. 2022).

1.9.3.2 How a Pfam family is built

What made Pfam different from earlier collections was its use of two alignments for each family, and of profile HMMs to connect them (Sonnhammer et al. 1997):

1. **Seed alignment.** A curator assembles a small, high-quality multiple alignment of representative members of the family. Sequences are taken from a non-redundant database of reference proteomes (pfamseq) and the alignment is checked by hand. Candidate families come mainly from proteins of known structure in the PDB, and from proteins in complete proteomes that have no Pfam match yet.
2. **Profile HMM.** A profile HMM is built from the seed alignment with HMMER (`hmmbuild`) (Eddy 1998).
3. **Search.** The HMM is searched against the full sequence database (`hmmsearch`). With HMMER3 (Eddy 2011), this became roughly a hundred times faster than before, which made frequent updates possible.
4. **Full alignment.** All sequences scoring above the family's threshold are aligned to the HMM. This gives the *full alignment*, which can contain millions of sequences.

The key advantage of this design is that only the small seed alignment has to be maintained by hand. When the sequence database grows, the full alignment is regenerated automatically by searching again with the HMM.

1.9.3.3 Thresholds: GA, TC and NC

Each Pfam family has a manually set **gathering threshold (GA)**, a bit score above which a match is accepted as a member of the family. HMMER scores are given both for the whole sequence and for each individual domain match, and the GA has a value for both. The curator chooses the threshold to include as many true members as possible while excluding false positives. Pfam also records two derived values:

- the **trusted cutoff (TC)**, the lowest score of any sequence included in the family;
- the **noise cutoff (NC)**, the highest score of any sequence that was *not* included.

The gap between NC and TC indicates how cleanly the family is separated from the background. False positives are recognised mainly as overlaps: if a region matches two different families that are not known to be related, at least one of the matches is probably wrong. Thresholds are re-checked at each release to prevent such overlaps. In practice, using the curated thresholds (`--cut_ga` in HMMER) is usually more reliable than using a fixed E-value cutoff for all families.

1.9.3.4 Entry types, clans and domains of unknown function

Pfam entries are classified as *families*, *domains*, *repeats* or *motifs*. "Family" is the default and just means that the members are related. "Domain" means an autonomous structural unit that can occur in different protein contexts. "Repeat" denotes units that are not stable on their own and only form a structure as tandem copies. "Motif" denotes short units, usually outside globular domains.

As families grew and more structures were solved, it became clear that many Pfam families are themselves distantly related. Since 2005 such families have been grouped into **clans** (Finn et al. 2006): sets of families that are believed to share a common evolutionary origin, based on structural similarity, functional similarity, and HMM–HMM comparison (Section 1.6). Pfam curators also use the ECOD structural classification (below) to identify clan relationships. A clan corresponds roughly to a superfamily in the structural classifications, but the two are defined on different units and need not coincide (see the table at the end of this section). Families in the same clan are not allowed to overlap in the same sequence region, so the best-scoring member family is reported.

A substantial part of Pfam consists of **domains of unknown function (DUFs)**: families that are clearly conserved across species but whose function has not been determined. They are named DUF followed by a number and are renamed when a function is found. Since around 2010 more than a fifth of all Pfam families have been DUFs, and their share has grown as new genomes are sequenced. Many DUFs are good targets for experimental work.

1.9.3.5 Pfam-A and Pfam-B

The curated families described above form **Pfam-A**. Since they did not cover all proteins, Pfam also provided **Pfam-B**, a large set of automatically generated families from sequence clustering. Pfam-B families were of lower quality, but could give a hint when no Pfam-A family matched. Pfam-B was discontinued with release 28.0 and later reintroduced (release 33.1), now generated with MMseqs2 clustering (Section 1.4).

1.9.3.6 Community curation through Wikipedia

Writing functional descriptions for thousands of families is more than a small curation team can do. From release 26.0, Pfam (together with its sister database Rfam for RNA families) therefore linked family descriptions to Wikipedia articles. Existing articles were connected to the corresponding families, and the community was invited to write new ones. Edits are reviewed by Pfam curators before they appear on the Pfam pages. Many Wikipedia articles on protein domains therefore originate from, and are maintained together with, Pfam.

1.9.4 Pfam-N: deep learning for domain annotation

Profile HMMs are sensitive, but they still miss many sequences that do belong to a family, especially in poorly studied lineages and metagenomes. In 2022, Maxwell Bileschi, Lucy Colwell and colleagues at Google Research showed that deep neural networks can be trained to classify protein sequences into Pfam families directly (Bileschi et al. 2022). Their model, **ProtCNN**, is a convolutional neural network (Chapter 12) trained on Pfam seed alignments. It takes a protein domain sequence as input and outputs the probability that it belongs to each Pfam family. An ensemble of such networks (**ProtENN**) was then combined with HMMER.

Two features of this work are important. First, the networks were evaluated on a split of the data in which test sequences had low sequence identity to the training sequences. This tests the ability to

detect remote homologues rather than near-duplicates, a point that applies to all machine-learning methods in bioinformatics (Chapter 12). On such a split, the deep models made fewer classification errors than profile HMMs, and the combination of the two was better than either. Second, the models were used to annotate UniProt sequences that had no Pfam match. This added several million new domain annotations and raised Pfam's coverage of UniProtKB by roughly a tenth. These predictions are distributed as **Pfam-N** ("N" for neural).

Pfam-N annotations are kept separate from the HMM-based ones, because they come with different evidence and different error properties. Curators also use them to extend the seed alignments of the standard families. This is part of a broader shift: recent Pfam releases also use AlphaFold-predicted structures to check and correct family boundaries, and to identify new families and clan relationships (Paysan-Lafosse et al. 2024).

1.9.5 InterPro

Pfam is one of several "signature" databases, each modelling families, domains or sites in its own way: regular expressions (PROSITE patterns), position-specific scoring matrices (CDD, PROSITE profiles), fingerprints (PRINTS) and profile HMMs (Pfam, SMART, TIGRFAMs/NCBIfam, PANTHER, CATH-Gene3D, SUPERFAMILY, PIRSF, HAMAP). **InterPro**, at EMBL-EBI, integrates these member databases into a single resource (Paysan-Lafosse et al. 2022). Signatures from different databases that describe the same family, domain or site are merged into one InterPro entry, with a common name, description and Gene Ontology terms. Entries are linked in parent–child relationships (for example, a family and its subfamilies) and classified by type (family, domain, repeat, site, homologous superfamily). InterPro precomputes matches for all of UniProtKB; more than 80% of its sequences have at least one match (Blum et al. 2020; Paysan-Lafosse et al. 2022). The InterProScan software lets users run all member databases on their own sequences. InterPro is now the standard starting point for annotating a new protein, and the databases are discussed further in Chapter 2.

1.9.6 Structural classification of domains

Because structure among homologues tends to be conserved much longer than sequence, many evolutionary relationships between domains can only be seen at the structural level. Several databases therefore classify domains of known three-dimensional structure. They are only introduced here; structure itself is the topic of Chapter 9.

1.9.6.1 SCOP

The **Structural Classification of Proteins (SCOP)** was created in 1994 by Alexey Murzin, Steven Brenner, Tim Hubbard and Cyrus Chothia at the MRC Laboratory of Molecular Biology and the Centre for Protein Engineering in Cambridge (Murzin et al. 1995). SCOP is largely manual: the classification was based on expert inspection of structures, supported by sequence comparison. Its unit is the structural domain, and its main levels are:

- **Class:** the overall secondary-structure content (all-α, all-β, α/β, α+β and some special classes).
- **Fold:** domains with the same major secondary structures in the same arrangement and with the same topological connections. A shared fold does *not* by itself imply common ancestry; similar folds can arise by convergence because only a limited number of arrangements are physically favourable.
- **Superfamily:** domains in the same fold that have low sequence identity but whose structural and functional features together indicate a probable common ancestor.

- **Family:** domains with clear sequence similarity (typically more than 30% identity), or lower identity combined with very similar structure and function. These are closely related homologues.

The human trypsin domain, for example, is classified as all-β class, trypsin-like serine protease fold, trypsin-like serine protease superfamily, eukaryotic protease family. The original SCOP was updated until version 1.75 (2009). It is continued as **SCOPe** at Berkeley (Fox et al. 2013), which adds new structures in a largely automated but compatible way, and as **SCOP2** from the Cambridge group (Andreeva et al. 2019), which replaces the strict tree with a more flexible network of relationships.

1.9.6.2 CATH

CATH was created in the mid-1990s by Christine Orengo, Janet Thornton and colleagues at University College London, and is still developed by the Orengo group (Sillitoe et al. 2020). Protein chains from the PDB are "chopped" into domains by a combination of algorithms and curation, and each domain is classified in four main levels, which give the database its name:

- **Class (C):** secondary-structure content (mainly α, mainly β, α–β, few secondary structures).
- **Architecture (A):** the overall arrangement of secondary structures in space, regardless of how they are connected (e.g. sandwich, barrel, three-layer sandwich).
- **Topology (T):** the fold, i.e. the arrangement *and* connectivity of the secondary structures.
- **Homologous superfamily (H):** domains with good evidence of a common ancestor.

Below the H level, CATH groups domains into sequence clusters and *functional families* (FunFams). Compared with SCOP, CATH relies more on automatic structure comparison, has an explicit architecture level, and does not use the α+β class. Through its sister resource **Gene3D**, CATH superfamilies are represented by HMMs that are used to assign domains to sequences without known structure, and Gene3D is one of the InterPro member databases.

1.9.6.3 ECOD

ECOD (Evolutionary Classification of Protein Domains), from Nick Grishin's group at UT Southwestern, puts evolutionary relationships first (Cheng et al. 2014). Its levels are *architecture* (A), *possible homology* (X), *homology* (H), *topology* (T) and *family* (F). The X level is the most distinctive: it groups domains for which homology is plausible but not established, instead of forcing a yes-or-no decision. Higher levels are defined by evidence of common ancestry rather than by shape alone, so domains with different topologies can still be placed in the same homology group if other evidence supports it. ECOD is updated weekly from the PDB, and its families map closely onto Pfam families, which is why Pfam uses it when defining clans.

1.9.6.4 TED: domains across the AlphaFold Database

Until recently, structural classifications covered only the roughly 200,000 experimentally determined structures in the PDB. The AlphaFold Protein Structure Database (Varadi et al. 2021) changed this by providing predicted structures for more than 200 million UniProt sequences. **The Encyclopedia of Domains (TED)** (Lau et al. 2024), from the Orengo and Jones groups at UCL together with others, used these models to segment and classify domains at the scale of the whole protein universe.

Chopping that many structures by hand is impossible. TED therefore combined three deep-learning domain-segmentation methods (Chainsaw, Merizo and UniDoc) and accepted domain boundaries on which they agreed. Applied to about 188 million unique AlphaFold models, this identified nearly 365 million domains. Using structure comparison (including Foldseek, Section 1.4), about 194 million of

these could be assigned to known CATH superfamilies, and a further 46 million matched known CATH topologies. The rest have structures not previously classified in CATH, and include many candidate new folds. TED shows both how much of protein space is covered by known folds, and how much remains unexplored.

1.9.6.5 Structure classifications and homology

The structural hierarchies connect directly to the definition of homology in 1. Homology means common ancestry; it is not a degree of similarity. Placing two domains in the same **superfamily** (SCOP, CATH) or **homology group** (ECOD) is a statement that there is enough evidence (structural similarity together with shared functional details, unusual structural features or weak sequence similarity) to conclude that they are homologous, even when their sequence identity is too low for sequence methods to detect it. Placing them only in the same **fold** or **topology** says that they have similar architectures, but leaves open whether this reflects common ancestry or convergence. The TIM barrel is the standard example: it is found in many enzyme families catalysing unrelated reactions, and it is still debated whether all TIM barrels share a common ancestor. The structure databases, in other words, help us infer homology, but similarity of structure, like similarity of sequence, is evidence for homology and not the same thing.

The same words have different meanings in different databases, which is a frequent source of confusion. Table 1.1 compares them.

Table 1.1: Domain-classification terms in different databases.

Term	Database	Meaning
Domain	SCOP, CATH, ECOD, TED	A structural unit, cut out of a structure (experimental or predicted).
	Pfam, InterPro	A region matched by a sequence model; in Pfam also one *entry type*, for units that occur in different protein contexts.
Family	SCOP	Domains with clear sequence similarity (often above 30% identity) or very similar structure and function.
	Pfam	The default entry type: any set of related sequences modelled by one HMM, which may be a domain, part of a domain or a whole protein.
	ECOD (F), CATH (FunFam)	ECOD: sequence-based families, close to Pfam families. CATH FunFams: functionally coherent subgroups within a superfamily.
Superfamily / homology group	SCOP, CATH (H), ECOD (H)	Domains judged to share a common ancestor from structural, functional and weak sequence evidence.
Possible homology	ECOD (X)	Homology plausible but not established.

Term	Database	Meaning
Clan	Pfam	Sequence families believed to share a common ancestor (from structure, function and HMM–HMM comparison). Roughly a superfamily, but built from Pfam families, so its boundaries and membership can differ.
Fold / topology	SCOP (fold), CATH (T)	Same secondary structures in the same arrangement and connectivity; *no* claim of common ancestry. In ECOD, T is instead a level *below* the homology group.

See also: further explanation in Section D.1.7; related Wikipedia articles in Section E.1.7.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Protein domain, Pfam, Structural Classification of Proteins database, CATH database and InterPro (CC BY-SA 4.0, retrieved 2 October 2026).

1.10 Protein language models

Every method in the previous sections compares sequences residue by residue, with scores that come from explicit models of evolution: substitution matrices, profiles, HMMs. Since about 2019 a different idea has spread from natural-language processing into biology. A large neural network is trained to *predict amino acids from their sequence context* on hundreds of millions of natural protein sequences, without any labels. To do this well, the network learns, implicitly, which substitutions are tolerated at which positions, which residues tend to be in contact and which sequence patterns belong to which families. Its internal representations, the *embeddings*, can then be used for prediction tasks and, more recently, to search for homologues. These networks are called **protein language models** (PLMs). This section describes where they come from, what they learn, and how they are being used for sequence searching and the detection of remote evolutionary relationships. The machine learning background is covered in Chapter 12.

1.10.1 From words to residues

1.10.1.1 The distributional hypothesis

In linguistics, the idea that "a word is characterised by the company it keeps" goes back to John Rupert Firth in 1957. This *distributional hypothesis* says that words occurring in similar contexts tend to have similar meanings. In natural-language processing it led to **word embeddings**: each word is represented by a real-valued vector, learned so that words that appear in similar contexts end up close together in the vector space. Early approaches reduced the dimension of word co-occurrence matrices, for example by latent semantic analysis in the late 1980s. From about 2000, neural probabilistic language models

learned distributed representations directly, and word2vec (2013) made this cheap enough to train on very large text collections.

A protein sequence can be read in the same way. The *tokens* are the twenty amino acids (plus a few special symbols), a protein is a "sentence", and the "corpus" is a sequence database such as UniProt or UniRef (Chapter 2). The analogy is useful but imperfect. The protein alphabet is tiny compared with a natural vocabulary, so a single residue carries little meaning on its own. What a residue "means" depends almost entirely on its context: the same leucine may be buried in a helix in one protein and exposed in a loop in another. A useful protein model therefore needs *contextual* embeddings, where the representation of a residue depends on the whole sequence around it. The evolutionary interpretation of the distributional hypothesis is also clear. The sequences in the corpus are the survivors of natural selection, so the patterns of co-occurring residues reflect structural and functional constraints. These are the same constraints that give rise to conserved columns in a multiple alignment (Section 1.6) and to correlated mutations between residues in contact (Chapter 9).

1.10.1.2 Language modelling objectives

A *language model* assigns probabilities to sequences. Two training objectives dominate.

- **Autoregressive (causal) language modelling.** The model predicts each residue from the residues before it, so that the probability of a sequence $x = x_1 \dots x_L$ factorises as

$$p(x) = \prod_{i=1}^{L} p(x_i \mid x_1, \dots, x_{i-1}).$$

  This is the objective of the GPT family in natural language. Such models can also generate new sequences, one residue at a time.
- **Masked language modelling.** This is the objective popularised by BERT (2018). A random subset M of positions (typically about 15%) is hidden, by replacing the residues with a mask token, and the model is trained to recover them from the remaining, unmasked context on *both* sides. The loss for one sequence is

$$\mathscr{L} = -\sum_{i \in M} \log p\left(x_i \mid x_{\setminus M}\right),$$

  where $x_{\setminus M}$ denotes the sequence with the masked positions hidden. The model never sees labels such as structure or function. This is *self-supervised* learning: the training signal is the sequence itself.

How well a model has learned is often summarised by its *perplexity*, the exponential of the average negative log-likelihood per residue. A model that knows nothing has a perplexity of about 20 (a uniform guess over the amino acids). Lower values mean the model predicts the hidden residues better.

1.10.1.3 Recurrent networks and transformers

The first protein language models used recurrent neural networks, in particular long short-term memory (LSTM) networks. These read the sequence one position at a time and carry information forward in a hidden state. They are hard to parallelise, and information from distant positions is easily lost.

The **transformer** (Vaswani et al. 2017) removed the recurrence. Each token is first converted into a vector by an embedding layer. Positional information is added, because attention by itself ignores order. Each layer then updates every token's vector by *self-attention*, a weighted combination of all the other

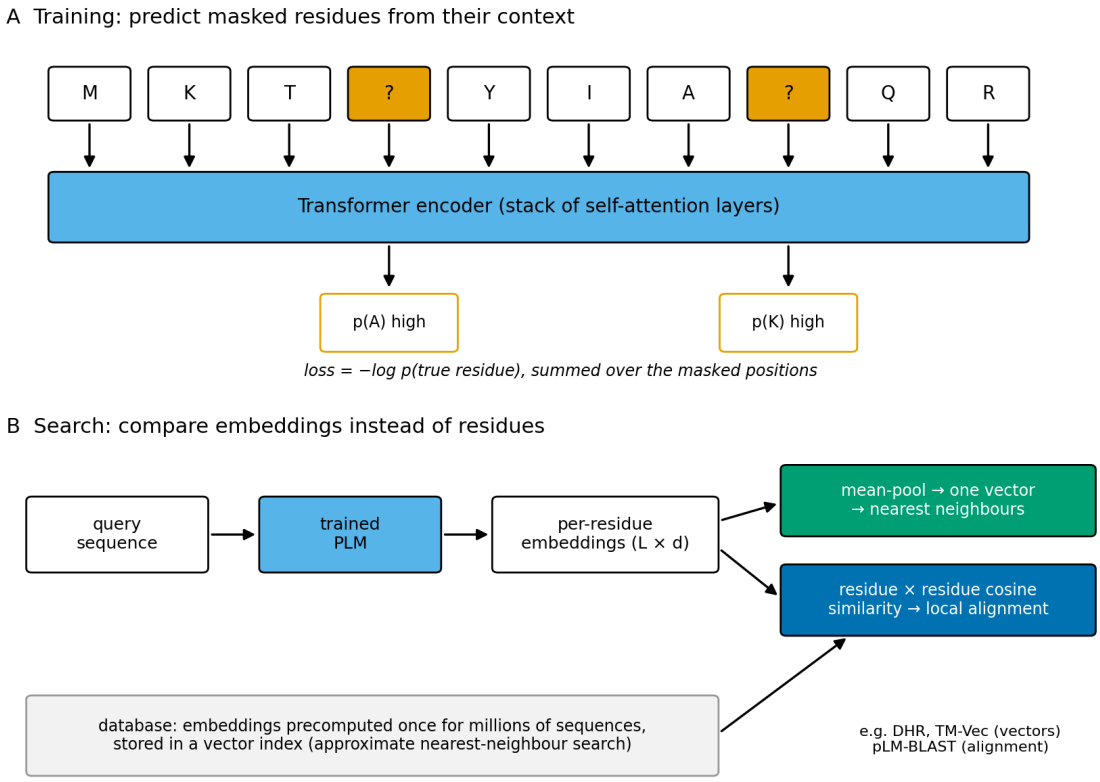


Figure 1.11: Protein language models. (A) In masked language modelling, some residues are hidden and the network is trained to predict them from the rest of the sequence. (B) After training, the per-residue embeddings can be compared directly, either as one pooled vector per protein (nearest-neighbour search) or residue by residue (alignment of embeddings).

tokens in the sequence. For a sequence represented as a matrix of token vectors, three learned linear maps give queries Q, keys K and values V, and attention is computed as

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

where d_k is the dimension of the keys. The matrix $\mathrm{softmax}(QK^{\top}/\sqrt{d_k})$ is an $L \times L$ table that says how much each position "looks at" every other position. Several such *attention heads* run in parallel in each layer, and many layers are stacked.

Two properties make transformers attractive for proteins. All positions are processed in parallel, so very large models can be trained on GPUs. And any position can interact directly with any other, which suits proteins, where residues far apart in sequence are often close in space. The price is that the cost of attention grows quadratically with sequence length. Most protein models are therefore trained on sequences cropped to around 1,000 residues.

1.10.2 A short history of protein language models

1.10.2.1 Early embeddings

Representing biological sequences with learned vectors predates the current models. In 2015 Asgari and Mofrad adapted word2vec to proteins ("ProtVec") (Asgari and Mofrad 2015), treating overlapping amino-acid 3-mers as words. These context-free embeddings carried some information about family membership and physico-chemical properties. However, they ignore the order and context of the 3-mers and were soon superseded.

1.10.2.2 Recurrent models: UniRep and SeqVec (2019)

The first models that learned contextual representations from large sequence collections appeared in 2019. **UniRep** (Alley et al. 2019) was a multiplicative LSTM trained autoregressively, to predict the next amino acid, on about 24 million UniRef50 sequences. Averaging its hidden states over the sequence gave a fixed-length vector per protein. Simple regression models on top of this vector could predict stability and the effects of mutations, and the authors used it to guide protein engineering. **SeqVec** (Heinzinger et al. 2019) adapted ELMo, a bidirectional LSTM language model from natural-language processing, to UniRef50. Per-residue embeddings from SeqVec, used as input to small supervised networks, predicted secondary structure and disorder. Per-protein embeddings predicted subcellular localisation and membrane association. These predictions approached, but did not reach, methods that used evolutionary information from multiple sequence alignments. They were, however, much faster, because no database search was needed.

Around the same time, the TAPE benchmark (Rao and colleagues, 2019) compared several architectures on a common set of tasks, such as secondary structure, contacts, remote homology, fluorescence and stability. It showed that self-supervised pre-training helps on almost all of them. On the structural tasks, however, it still did not reach features derived from alignments.

1.10.2.3 Transformers at scale: ESM-1b and ProtTrans (2020–2021)

The next step was to train much larger transformers on much more data. **ESM-1b** (Rives et al. 2021), from the group of Alexander Rives at Facebook AI Research, was a 33-layer transformer with about 650 million parameters. It was trained with masked language modelling on UniRef50, which the authors

drew from a collection of about 250 million sequences. The paper analysed what the model had learned without supervision:

- the embeddings of the amino acids themselves clustered by physico-chemical properties (hydrophobic, aromatic, charged and so on);
- per-protein embeddings separated orthologous groups and reflected phylogenetic relationships;
- per-residue embeddings contained information on secondary structure and on which residues are in contact;
- these signals improved as the model's perplexity decreased.

Shortly afterwards, Rao and colleagues showed that the attention maps of such models predict residue–residue contacts directly (Rao et al. 2020). A simple logistic regression on the attention weights recovered contacts with accuracy comparable to coevolution methods based on a full multiple alignment. The model had, in effect, learned something similar to the coupling analysis of Chapter 9, but from single sequences.

ProtTrans (Elnaggar et al. 2022), from the group of Burkhard Rost in Munich, trained several transformer architectures from natural-language processing on UniRef and on the metagenomic Big Fantastic Database (BFD), with billions of sequences in total. The models included ProtBERT, an encoder trained by masked language modelling, and ProtT5, an encoder–decoder model. The encoder of ProtT5-XL became one of the most widely used sources of embeddings. Using only its embeddings, without alignments, the authors predicted secondary structure about as well as the best alignment-based methods of the time. They also predicted localisation and membrane association well. A review by Bepler and Berger (2021) summarises this period and discusses how evolution, structure and function are encoded in the learned representations.

1.10.2.4 Structure from a language model: ESM-2 and ESMFold (2022–2023)

ESM-2 (Lin et al. 2023) scaled the masked-language-model approach to a series of models from 8 million to 15 billion parameters, trained on UniRef. As the models grew, their perplexity fell and the contact information in their attention maps increased. **ESMFold** attaches a structure module, similar to that of AlphaFold2 (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021), to the ESM-2 language model. It predicts atomic structures from a *single sequence*, without searching for homologues or building an alignment. ESMFold is less accurate than AlphaFold2 on average, especially for proteins with few homologues. Because it skips the alignment step, however, it is roughly an order of magnitude faster. The authors used it to predict more than 600 million structures from metagenomic sequences (the ESM Metagenomic Atlas). An important observation was that ESMFold's accuracy is closely related to the language model's perplexity on the sequence. The model works well when it "understands" a protein, which in practice usually means that related sequences were well represented in its training data.

1.10.2.5 Generative and multimodal models (2024–)

The most recent models go beyond sequence. **ESM3** (Hayes et al. 2025) is a generative masked language model that reasons over three *tracks* at once: sequence, structure and function. Structure is encoded as discrete tokens, and function as keyword annotations. Any combination can be given as a prompt and the rest generated. The largest version has about 98 billion parameters. As a demonstration, the authors generated a functional green fluorescent protein, esmGFP, that is only distantly related to known fluorescent proteins. This illustrates a broader trend: language models are increasingly used to *design*

sequences (Chapter 9), not just to analyse them. Other models in this family include autoregressive models trained to generate protein families, and "bilingual" models that translate between amino-acid sequence and structural alphabets.

1.10.3 What do embeddings capture?

It is worth being precise about what a PLM provides. For a protein of length L, the last layer of the network (or a chosen intermediate layer) gives one vector of dimension d per residue, an $L \times d$ matrix. Typically d is between about 1,000 and 5,000. This **per-residue embedding** is used for tasks that need a prediction for each residue, such as secondary structure, disorder, binding sites, transmembrane segments or signal peptides (Chapter 10). Averaging the vectors over all positions (*mean pooling*) gives a **per-protein embedding** of fixed size d. This is used for tasks about the whole protein, such as localisation, enzyme class or family membership, and for fast comparison of proteins.

"The embedding of a protein" is therefore not a uniquely defined object. It depends on the model, on the layer it is taken from (intermediate layers are sometimes better than the last), and, for a per-protein vector, on the *pooling*: mean pooling, the vector of a special start token, maximum pooling, or a learned (attention-based) pooling. Long sequences may also be cropped or split into chunks, which changes the result. Pooling discards the order of the residues, so two proteins with similar composition and local environments but different domain order can get similar per-protein vectors. Embeddings are therefore only comparable when they were computed with the same model, layer and pooling.

Several lines of evidence show what these representations contain:

- **Structure.** Secondary structure and solvent accessibility can be read from per-residue embeddings with small supervised models. Contacts can be read from attention maps, and full 3D structures can be built from the representations (ESMFold).
- **Function.** Per-protein embeddings cluster by enzyme class, Gene Ontology terms and subcellular localisation. Simple nearest-neighbour annotation in embedding space ("embedding-based annotation transfer") often works where sequence similarity is too low for BLAST.
- **Evolution.** The likelihood a masked model assigns to a mutation correlates with the measured fitness effect of that mutation in deep mutational scanning experiments. This makes PLMs useful as unsupervised predictors of variant effects (Meier et al. 2021) (Chapter 10).
- **Family membership.** Proteins from the same family or superfamily are often close in embedding space, even when sequence identity is in the twilight zone (1). This is an empirical ability to recover family and structural relationships, not a direct inference of ancestry. Closeness in embedding space can also reflect shared structure class, composition or function, so on its own it is not evidence of common ancestry in the sense of a significant alignment score.

A common interpretation is that a large PLM acts as a *compressed multiple sequence alignment*. During training it sees many members of each family. The weights then store the conservation and covariation patterns that a profile or coevolution model would extract from an explicit alignment, and the model "retrieves" this information when it sees a new member of the family. This view explains both the strengths and the main weakness of PLMs. They are good for proteins from large, well-sampled families. For *orphan* proteins, or families with few sequences in the training data, they know little more than what is in the query itself.

1.10.4 Using language models to search for homologues

Classic homology search compares residues with a substitution matrix (Section 1.3) or a profile (Section 1.6). Embeddings offer two new possibilities. First, two proteins can be compared through their embeddings instead of their residues. Second, because an embedding summarises a sequence in context,

it may carry information about structure and family that the raw sequence only reveals through an alignment with many homologues. Several types of methods have appeared since 2021.

1.10.4.1 Embedding nearest-neighbour search

The simplest approach compares per-protein embeddings, for example with the cosine similarity

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

The embeddings for a whole database are computed once and stored in a *vector index*. Specialised libraries (for example FAISS (Johnson et al. 2021)) can then find the approximate nearest neighbours of a query among hundreds of millions of vectors in milliseconds, using clustering and quantisation of the vectors. Embedding the query is a single pass through the network, so the search itself costs almost nothing compared with dynamic programming against every database sequence.

Embeddings from a model trained only by masked language modelling are not optimised for this purpose. Better results come from *fine-tuning* the encoder with *contrastive learning*, so that homologous pairs end up close together and unrelated pairs far apart. This is the idea behind **dense retrieval** in information retrieval, where documents and queries are encoded into the same vector space. The **Dense Homolog Retriever** (DHR) (Hong et al. 2024) fine-tunes a protein language model contrastively, using homologous pairs taken from family alignments. It retrieves homologues with sensitivity comparable to or better than profile methods on the authors' benchmarks, at much higher speed. The authors also used it to build multiple sequence alignments for structure prediction quickly. This links PLM-based search to the alignment-building step of AlphaFold2 (Chapter 9).

1.10.4.2 Predicting structural similarity from sequence

Because structure among homologues tends to be conserved longer than sequence, structural similarity is the most sensitive sign of remote homology (1). It normally requires structures, which led to Foldseek and its 3Di alphabet (van Kempen et al. 2023) (Section 1.4). **TM-Vec** (Hamamsy et al. 2023) instead trains a network on top of PLM embeddings to predict the *TM-score* between two proteins, a measure of structural similarity (Chapter 9), directly from their sequences. The network is trained on pairs of proteins with known structures, so that the cosine similarity of the two output vectors approximates their TM-score. A database is again indexed once, and a query returns the proteins predicted to have similar structures, even when sequence identity is very low. The companion method DeepBLAST predicts a structural alignment between the two sequences.

A related idea uses language models to predict the 3Di structural alphabet directly from sequence (for example ProstT5 (Heinzinger et al. 2024), a "bilingual" model derived from ProtT5). This allows Foldseek-style structural search for proteins without any predicted structure.

1.10.4.3 Aligning embeddings

Pooling an $L \times d$ embedding into a single vector throws away the positional information needed for an alignment. An alternative keeps the per-residue embeddings and uses them in place of a substitution matrix. **pLM-BLAST** (Kaminski et al. 2023) computes the cosine similarity between every residue embedding of the query and every residue embedding of the target. This gives an $L_1 \times L_2$ similarity matrix specific to that pair of sequences, which is then used in a Smith–Waterman-like local alignment (Section 1.3). Since the embedding of a residue depends on its context, the matrix effectively contains a position-specific substitution score for each pair, much like a profile–profile comparison, but obtained from two single sequences. On benchmarks of distant homology, pLM-BLAST detects relationships

that BLAST misses and performs comparably to HMM–HMM comparison with HHsearch (Section 1.6), without building alignments for either sequence. Methods of this type also show where in two sequences the similarity lies, which pooled embeddings cannot.

1.10.4.4 Combined pipelines

Practical tools combine these ideas to balance speed and sensitivity, in the same way that BLAST and MMseqs2 combine a fast prefilter with a slower alignment step (Section 1.4). **PLMSearch** (Liu et al. 2024) first filters candidate pairs, using predicted family membership and embedding similarity. It then scores the remaining pairs with a model that predicts structural similarity from ESM embeddings, and finally aligns the hits with an embedding-based alignment method. The authors reported sensitivity for remote homologues similar to structure-based search, at a speed close to that of MMseqs2.

1.10.4.5 Deep learning for family annotation

Closely related to search is the assignment of sequences to known families. **ProtCNN** (Bileschi et al. 2022) is a deep convolutional network trained to classify sequence domains into Pfam families. It is not a transformer language model, but it rests on the same idea of learned sequence representations. Combined with profile HMMs, it was used to create *Pfam-N*, which added millions of new family assignments to UniProt sequences that the Pfam HMMs had missed. This is discussed in Section 1.9.

1.10.5 How good are they?

Remote-homology methods are usually evaluated on benchmarks derived from structural classifications such as SCOP or CATH (Section 1.9). Pairs in the same superfamily but in different families count as true remote homologues, and pairs in different folds count as non-homologues. Methods are compared by how many true pairs they rank above the first false one, or by the area under ROC-type curves. On such benchmarks, embedding-based methods are often more sensitive than BLAST and MMseqs2 in the twilight zone, and in several studies match or exceed HMMER and HHblits (Eddy 2011; Remmert et al. 2011). They usually remain below full structural comparison when structures are available.

These results have to be read carefully, for reasons that are directly relevant to the exam's emphasis on understanding the limitations of methods.

- **Leakage between training and test data.** A language model is trained on essentially all known sequences, including the test proteins and their relatives. A supervised model on top of it may be trained on pairs from the same families as the benchmark. Unless training and test sets are separated by family or superfamily, not just by sequence identity, reported performance is likely to be optimistic (the same issue as in any supervised learning in biology; 4).
- **No well-founded statistics.** The great strength of BLAST, HMMER and MMseqs2 is that their scores come with E-values derived from a model of random alignments (Section 1.4). A cosine similarity or a predicted TM-score has no such theory behind it. Thresholds are set empirically, and the false-positive rate on a new database is hard to predict. Some methods calibrate their scores against shuffled or unrelated sequences, but this is less mature than for classic tools.
- **Bias towards well-sampled families.** As discussed above, PLMs work best where training data are rich. Orphan proteins, rapidly evolving families, viral proteins and *de novo* designed proteins are often poorly represented, and these are precisely the cases where a sensitive remote-homology method is needed most.

- **Cost.** Searching an index is fast, but computing embeddings for hundreds of millions of sequences with a large model needs GPUs and a lot of storage. Each per-residue embedding is thousands of numbers per residue. The indexes must also be rebuilt when the model changes.
- **Interpretability.** An HMMER hit can be explained by the alignment and the profile columns that match. An embedding similarity is much harder to explain, although methods that align embeddings (pLM-BLAST) recover some of this.

1.10.6 Outlook

PLMs have not replaced classic sequence searching. They have added a new layer to it. Profile HMMs and fast *k*-mer-based tools remain the standard for routine annotation, because they are well understood, their statistics are reliable, and they run on ordinary computers. Language models are most useful at the edge of what these methods can detect: in the twilight zone, for building alignments quickly, for annotating metagenomic sequences at scale, and as cheap stand-ins for structural comparison. The classic and new approaches are also merging. Profile and alignment information is used to train language models, and language models are used to produce substitution scores, structural alphabets and alignments that feed into dynamic programming. It is therefore important to understand both: what the older methods assume, and why a model trained only to fill in missing amino acids ends up with representations from which evolutionary, structural and functional relationships can be recovered.

See also: further explanation in Section D.1.8; related Wikipedia articles in Section E.1.8.

> **Sources**
>
> The background on word embeddings, the distributional hypothesis and the transformer architecture is adapted from the Wikipedia articles Word embedding, Transformer (deep learning) and Large language model (CC BY-SA 4.0, retrieved 2 October 2026). There is no Wikipedia article on protein language models. The rest of the section is based on the papers cited.

1.11 Benchmarking search methods

Each method in this chapter claims to be more sensitive, faster or more accurate than its predecessors. At the exam you should be able to say how such claims are tested, and what can go wrong. The central problem is that we need to know the right answer: which pairs of sequences are truly homologous and which are not. That knowledge must come from somewhere other than the method being tested.

1.11.1 Where does the truth come from?

The most widely used source of truth is a **structural classification** (Section 1.9). Because structure among homologues tends to be conserved much longer than sequence, two domains classified in the same SCOP or CATH superfamily are taken to be homologous, even if they have no detectable sequence similarity. Domains in different folds are taken to be unrelated. Pairs in the same fold but different superfamilies are ambiguous (they may be distant homologues or analogues) and are usually excluded. These labels are an *operational* gold standard: a convenient, independent definition of right and wrong answers, not a complete record of evolutionary history. Early benchmarks used curated sequence superfamilies instead (Pearson 1995). The SCOP-based benchmark of Brenner et al. (1998) became the template for later work. It showed that pairwise methods found only a small fraction of the homologous pairs with less than about 30% identity, and that well-computed E-values were reliable indicators of significance. Park et al. (1998) used the same approach to show that profile methods find about three

times as many remote homologues, and Lindahl and Elofsson (2000) measured performance separately at the family, superfamily and fold levels. Their results showed that methods which work well for close relationships can fail almost completely at the fold level, so a single overall number can hide large differences.

Other reference sets are used for other tasks: structural superpositions for multiple alignments (BAliBASE, Thompson et al. (1999)), and blind prediction experiments for structure and function (CASP and CAFA; Chapter 9 and Chapter 10).

1.11.2 What is measured?

A search method returns a ranked list of hits with scores or E-values. Performance is described by how true and false hits are ordered in that list.

- **ROC curves.** The *receiver operating characteristic* curve plots the number (or fraction) of true positives against false positives as the score threshold is lowered. The area under the curve (AUC) summarises it. In a database search almost all pairs are negatives, and only the top of the list is ever examined. The **ROC$_n$** score is therefore often used. It is the area under the curve up to the nth false positive, for example ROC$_1$ or ROC$_{50}$ (Gribskov and Robinson 1996).
- **Coverage versus errors per query.** Brenner et al. (1998) plotted the fraction of true relationships found (coverage) against the number of false positives per query. This shows directly how many homologues are found at an acceptable error rate.
- **Sensitivity up to the first false positive**, per query, averaged over all queries, is a common simple summary (used, for example, in the Foldseek and MMseqs2 papers; Section 1.4).
- **E-value calibration.** If E-values are correct, a search should produce on average E false positives with E-value at most E per query. Plotting the observed number of false positives against the reported E-value tests whether a method's statistics can be trusted. This is as important as sensitivity, because users choose thresholds based on E-values.
- **Speed**, in queries per hour or as a speed-up over BLAST, is always reported with a sensitivity measure. A faster method is only an improvement if it loses little sensitivity.

1.11.3 Pitfalls

Benchmarks are easy to design badly, and many published comparisons contradict each other for that reason.

- **Circularity.** If the reference set was built with one of the methods being compared, that method has an advantage. For example, Pfam families are defined by HMMER searches, so HMMER will look good on a benchmark based on Pfam membership.
- **Bias towards large families.** If every pair counts equally, a few very large superfamilies dominate the result. Results are usually weighted per query or per family, and redundant sequences are removed (for example, keeping only domains below 40% identity to each other).
- **"False" positives that are true.** Disagreement with a database classification is not automatically a biological false positive. Structural classifications are incomplete: relationships between superfamilies are still being discovered, and two superfamilies may later be merged or placed in the same ECOD homology group. Domain boundaries differ between SCOP, CATH and ECOD, and the schemes use different criteria for homology (Table 1.1). A hit counted as false may therefore be a real homologue classified in another superfamily, or a match to another domain in a multidomain protein. Ambiguous pairs should be excluded rather than counted as errors, and apparently false hits with very good scores are worth inspecting: some of them are discoveries.

- **Leakage in machine-learning methods.** A method that has been trained on sequences from the same families as the test set can appear to detect remote homologues while in fact recognising families it has seen. Training and test sets must be separated by superfamily or fold, not only by sequence identity. This is particularly important for deep-learning and language-model methods (Section 1.10), whose pre-training data may include the benchmark proteins themselves. Homology between training and test (or benchmark) sets is one of the most common reasons for inflated performance in bioinformatics, and it affects structure, function and interaction predictors as much as search methods; see 4.
- **Tuning on the test set.** Parameters chosen to maximise performance on a benchmark will overestimate how well a method works on new data. Independent test sets, or blind assessments such as CASP, avoid this.

When you read a claim that a new method "outperforms" existing ones, it is worth asking: what was the reference, how were true and false pairs defined, which measure was used, and could the method have seen the test data?

See also: further explanation in Section D.1.9; related Wikipedia articles in Section E.1.9.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia article Receiver operating characteristic (CC BY-SA 4.0, retrieved 2 October 2026).

1.12 Key concepts

A checklist for revision. You should be able to explain each point, including the algorithm or argument behind it, and its limitations.

- **Homology:** definition as common ancestry (qualitative, not a percentage); orthologues, paralogues (in-/out-paralogues), xenologues, ohnologues; analogy and convergence; why significant similarity implies homology but not the reverse; the twilight and midnight zones; why proteins are compared rather than DNA
- **Dynamic programming:** principle of optimality; Needleman–Wunsch and Smith–Waterman recursions and traceback; semi-global, overlap and banded alignment; linear, affine and convex gap penalties; Gotoh's three-matrix algorithm; time and space complexity, Hirschberg/Myers–Miller linear space
- **Substitution matrices:** log-odds scores, target and background frequencies, relative entropy; how PAM and BLOSUM are derived and how their numbering differs; choosing a matrix
- **Heuristic search:** FASTA (k-tuples, diagonals, banded optimisation); BLAST (neighbourhood words, threshold T, X-drop, two-hit method, gapped extension); low-complexity filtering; the BLAST family and translated searches; DIAMOND, MMseqs2 and Linclust; Foldseek and the 3Di alphabet
- **Statistics:** why local scores follow a Gumbel distribution; Karlin–Altschul $E = Kmne^{-\lambda S}$; bit scores, E- and P-values; effective lengths, gapped and composition-based statistics; why E-values depend on database size while the score does not; per-query E-values versus multiple testing over many queries; query and target coverage
- **From hit to annotation:** checking the query (gene model, fragments, low complexity); identity with its denominator, coverage of both sequences and domain architecture; why the best (or reciprocal best) hit is not necessarily the orthologue; escalating to profiles, structure or embeddings without lowering the evidence standard; annotation statements that match the evidence; recording database release, program version and parameters

- **Exact matching:** suffix trees and arrays; the Burrows–Wheeler transform, LF mapping, FM-index and backward search; why they suit read mapping
- **Multiple alignment:** no single correct objective (SP, consistency, probabilistic and structure-aware surrogates); sum-of-pairs score and NP-hardness; progressive alignment and "once a gap, always a gap"; ClustalW, Clustal Omega, T-Coffee (consistency), MUSCLE and MAFFT (iterative refinement); benchmarks; alignment uncertainty, MUSCLE5 ensembles and trimming as a modelling choice
- **Motifs and profiles:** consensus, PROSITE patterns, PWMs with pseudocounts, information content and logos; Gibbs sampling and MEME; Gribskov profiles; sequence weighting and priors; computing a PSSM with background-weighted pseudocounts; PSI-BLAST and profile corruption (how to recognise and limit it); PSSM vs. profile HMM vs. pair HMM
- **Hidden Markov models:** states, transitions, emissions; Viterbi, forward, backward, posterior decoding and Baum–Welch; profile HMM architecture (match/insert/delete, Plan 7), local and glocal modes; HMMER3 filter pipeline and forward-score E-values; per-sequence vs. per-domain scores, independent vs. conditional E-values, reporting vs. inclusion thresholds; HHsearch/HHblits (HMM–HMM comparison); sequence–sequence < sequence–profile < profile–profile
- **Domains:** structural, folding and evolutionary definitions; multidomain proteins and architectures; why domains complicate annotation transfer; Pfam (seed/full alignments, gathering thresholds, clans), Pfam-N, InterPro; SCOP, CATH, ECOD and TED, and fold vs. superfamily; database-specific meanings of family, superfamily and clan
- **RNA homology search:** why structured RNA homologues often conserve structure better than sequence; compensatory mutations and covariation; Nussinov and Zuker folding algorithms; SCFGs and their relation to HMMs (CYK, inside–outside); covariance models, Infernal and its HMM filters; Rfam; limitations (pseudoknots, unstructured ncRNAs)
- **Whole-genome alignment and conservation:** why genomes are not aligned like genes (rearrangements, duplications, repeats, soft-masking); BLASTZ/LASTZ with spaced seeds; chains and nets; reference-based (MULTIZ) vs. reference-free (Enredo–Pecan, Progressive Cactus) multiple genome alignment and the Zoonomia alignment; Alignathon and the lack of ground truth; neutral models; phastCons (phylo-HMM), phyloP (conservation and acceleration, human accelerated regions) and GERP (rejected substitutions); use for non-coding elements and variant interpretation; caveats (alignment errors, branch length needed, lineage-specific function, biased gene conversion)
- **Alignment-free comparison:** k-mer statistics; MinHash and the Mash distance; ANI and Kraken; when alignment-free methods work and when they fail
- **Protein language models:** masked and autoregressive objectives; transformers and attention; UniRep, SeqVec, ESM, ProtTrans, ESMFold, ESM3; per-residue and per-protein embeddings and pooling; embedding search, TM-Vec, pLM-BLAST, PLMSearch; limitations (leakage, lack of statistics, bias towards well-sampled families)
- **Benchmarking:** structural classifications as an operational ground truth (and why disagreement is not always an error); ROC, ROC$_n$, coverage vs. errors per query, E-value calibration; circularity, family-size bias and train–test leakage

1.13 Recommended reading

These papers are recommended for reading in full:

- Smith and Waterman (1981) — The local alignment algorithm, in three pages.
- Altschul et al. (1997) — Gapped BLAST and PSI-BLAST; still the standard reference for heuristic search and profiles.

- Eddy (2011) — HMMER3; profile HMM search made as fast as BLAST, with a clear account of the statistics.
- Steinegger and Söding (2017) — MMseqs2; how sensitive search scales to billions of sequences.
- van Kempen et al. (2023) — Foldseek; searching with structure instead of sequence.

Titles, abstracts, links and the full texts are in Section A.1; a longer list of further papers is in Section B.1.

1.14 Background reading (open access)

- **Sequence homology (Wikipedia)** — Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life.
  https://en.wikipedia.org/wiki/Sequence_homology
- **Sequence alignment (Wikipedia)** — In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.
  https://en.wikipedia.org/wiki/Sequence_alignment
- **Needleman–Wunsch algorithm (Wikipedia)** — The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970.
  https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm
- **Smith–Waterman algorithm (Wikipedia)** — The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences.
  https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm
- **Substitution matrix (Wikipedia)** — In bioinformatics and evolutionary biology, a substitution matrix describes the frequency at which a character in a nucleotide sequence or a protein sequence changes to other character states over evolutionary time.
  https://en.wikipedia.org/wiki/Substitution_matrix
- **BLOSUM (Wikipedia)** — In bioinformatics, the BLOSUM matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.
  https://en.wikipedia.org/wiki/BLOSUM
- **PAM (Wikipedia)** — A point accepted mutation — also known as a PAM — is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection. This definition does not include all point mutations in the DNA of an organism.
  https://en.wikipedia.org/wiki/Point_accepted_mutation
- **BLAST (Wikipedia)** — In bioinformatics, BLAST is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins, nucleotides of DNA and/or RNA sequences.
  https://en.wikipedia.org/wiki/BLAST_(biotechnology)
- **FASTA (Wikipedia)** — FASTA is a DNA and protein sequence alignment software package first described by David J. Lipman and William R. Pearson in 1985. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.
  https://en.wikipedia.org/wiki/FASTA
- **NCBI: The Statistics of Sequence Similarity Scores** — Stephen Altschul's tutorial on the statistics of local alignment scores: the extreme value distribution, E-values, P-values and bit

scores, as used by BLAST.
https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

- **Hidden Markov model (Wikipedia)** — In probability theory, a hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process. An HMM requires that there be an observable process whose outcomes depend on the outcomes of in a known way.
https://en.wikipedia.org/wiki/Hidden_Markov_model
- **Forward–backward algorithm (Wikipedia)** — The forward–backward algorithm is an inference algorithm for hidden Markov models which computes the posterior marginals of all hidden state variables given a sequence of observations/emissions , i.e. it computes, for all hidden state variables , the distribution . This inference task is usually called smoothing.
https://en.wikipedia.org/wiki/Forward%E2%80%93backward_algorithm
- **Additive smoothing ("pseudocounts") (Wikipedia)** — In statistics, additive smoothing, also called Laplace smoothing or Lidstone smoothing, is a technique used to smooth count data, eliminating issues caused by certain values having 0 occurrences.
https://en.wikipedia.org/wiki/Additive_smoothing
- **Multiple comparisons problem (Wikipedia)** — Multiple comparisons, multiplicity or multiple testing problem occurs when many statistical tests are performed on the same dataset. Each test has its own chance of a Type I error, so the overall probability of making at least one false positive increases as the number of tests grows.
https://en.wikipedia.org/wiki/Multiple_comparisons_problem
- **Family-wise error rate (Wikipedia)** — Family-wise error rate (FWER) is a term from statistics for the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests. FWER is a metric that quantifies the risk caused by multiple testing.
https://en.wikipedia.org/wiki/Family-wise_error_rate
- **Multiple sequence alignment (Wikipedia)** — Multiple sequence alignment (MSA) is the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. These alignments are used to infer evolutionary relationships via phylogenetic analysis and can highlight homologous features between sequences.
https://en.wikipedia.org/wiki/Multiple_sequence_alignment
- **A tutorial on Burrows–Wheeler indexing methods** — A step-by-step blog tutorial (The Grand Locus) on the Burrows–Wheeler transform, the FM-index and backward search, the data structures behind read mappers such as BWA and Bowtie.
http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

1.15 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

1.16 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Sequence Alignment and Search Foundations" and "Sequence Databases, BLAST, and Profile HMMs" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

1.16.1 Derive global and local alignment

Derive the dynamic-programming recurrences for global and local pairwise sequence alignment. Explain which biological questions each formulation answers, and why neither recurrence gives a uniquely true evolutionary alignment.

Extensions, solution outline and common errors

Extensions:

- Give the optimal-substructure argument from the three possible final columns of an alignment.
- Compare initialisation, termination and traceback for global, local and semi-global alignment.
- Derive the time and memory complexity for sequences of lengths n and m.
- How would you test whether the result for a protein pair with uncertain terminal homology depends on the alignment mode?

Solution outline:

1. Define F_{ij} as the optimal score of the prefixes $A_{1:i}$ and $B_{1:j}$; the final column must be a residue pair or a gap in one of the two sequences.
2. Write the recurrence with a diagonal, a vertical and a horizontal move, using a substitution score and a linear gap cost.
3. For global alignment, penalise leading gaps and trace back from F_{nm}. For local alignment, add zero to the maximum, initialise the first row and column to zero, start the traceback at the highest cell and stop at a zero.
4. State $O(nm)$ time and $O(nm)$ memory for the traceback matrix; distinguish computing the score alone in linear memory from reconstructing the alignment (Hirschberg, Myers–Miller).
5. Use global alignment for sequences expected to correspond over most of their lengths, local alignment for a shared domain or a fragment, and semi-global alignment when terminal gaps should be free.
6. Conclude that the optimum depends on the substitution scores, the gap model, the boundary conditions and any masking.

Common errors:

- Describing the optimum under one scoring system as the biological truth.
- Changing only the traceback, without the zero and the boundary conditions of local alignment.
- Aligning proteins with different domain architectures globally without considering alternatives.

Connections: Section 1.3; insertions and deletions in evolution (Chapter 3); alignments as uncertain input to phylogeny (Section 4.4); structure-based alignment (Section 9.6).

1.16.2 Interpret substitution scores and affine gaps

Explain why empirical substitution matrices and affine gap costs are used in protein alignment. Derive their computational representations, and identify the evolutionary approximations that they make.

Extensions, solution outline and common errors

Extensions:

- Derive the log-odds interpretation of $s(a, b)$.
- Contrast the construction and intended use of PAM and BLOSUM matrices.
- Explain why affine gaps require separate match and gap states (Gotoh's algorithm).
- Predict how changing g_o and g_e affects the number and length of gaps in an alignment.

Solution outline:

1. Compare the frequency q_{ab} of the pair in alignments of related sequences with the frequency $p_a p_b$ expected by chance, giving $s(a, b) = \lambda^{-1} \log[q_{ab}/(p_a p_b)]$.
2. Positive scores favour the model of related sequences, negative scores the background model, under the assumed frequencies.
3. PAM matrices extrapolate a Markov model of accepted mutations from close relatives; BLOSUM matrices are estimated directly from conserved ungapped blocks after clustering sequences above a given identity.
4. Larger PAM numbers, but smaller BLOSUM numbers, are meant for more divergent sequences; the best matrix still depends on the task.
5. Define $g(k) = g_o + (k-1)g_e$ and show that separate dynamic-programming states are needed to distinguish opening a gap from extending one.
6. Gap penalties are a convenient approximation to insertion and deletion processes, not a mechanistic model of them.

Common errors:

- Calling a conservative, positive-scoring substitution an identity.
- Claiming that one substitution matrix is best for all purposes.
- Copying gap parameters between programs without checking whether a gap of length one costs g_o or $g_o + g_e$.

Connections: Section 1.3; divergence and protein evolution (Section 3.5); substitution models in phylogenetics (Section 4.3.6); likelihood ratios and model validation (4).

1.16.3 Interpret a database-search result

A BLAST hit has $E = 10^{-20}$, 35% identity and 60% query coverage. What may be concluded, what remains unresolved, and what further information is needed?

Extensions, solution outline and common errors

Extensions:

- Define the E-value and distinguish it from the probability that the sequences are homologous.
- Explain how the size of the database affects the E-value for a fixed raw score.
- Compare raw score, bit score, percent identity and the coverage of both sequences.
- Design the checks needed before transferring a molecular function.

Solution outline:

1. Ask for the alignment length, target coverage, query and target lengths, the gaps, the

composition of the aligned region, the database release and the search parameters.
2. Define the E-value as the expected number of chance alignments with at least this score, under the null model and in a search space of this size.
3. Use $E \approx Kmne^{-\lambda S} = mn\,2^{-S'}$ to relate the raw score S, the bit score S', the effective lengths m and n, and the parameters K and λ.
4. After checking for low complexity and masking, the hit is strong evidence of local similarity; decide whether the aligned segment covers a whole protein or only a domain.
5. Homology is then the most likely explanation for the aligned region; orthology cannot be inferred from the rank of a hit.
6. Before assigning a detailed function, require conserved functional residues, a compatible domain architecture, evolutionary context and, preferably, experimental evidence.

Common errors:

- Reading 10^{-20} as the probability that the sequences are not homologous.
- Reporting percent identity without saying over how many positions.
- Equating the lowest E-value with orthology or with the same function.

Connections: Section 1.4; database composition and annotation errors (Section 2.2); homology and orthology (1, Section 3.4); function prediction (Section 10.4).

1.16.4 Compare exact and heuristic search

Explain why BLAST is faster than running Smith–Waterman against every database sequence, and describe the true relationships that a seed-and-extend heuristic can miss.

Extensions, solution outline and common errors

Extensions:

- Divide seed-and-extend into seeding, extension, gapped alignment and statistics.
- Identify the parameters that control the trade-off between sensitivity and speed.
- Compare the overall strategies of BLAST and MMseqs2.
- How would you compare two search programs fairly?

Solution outline:

1. Smith–Waterman fills the whole dynamic-programming matrix for every pair; BLAST first finds short high-scoring words and extends only around them.
2. Gapped alignment is computed only for a small set of candidates, so most database sequences are never aligned; the scores of the alignments that are computed mean the same as before.
3. Word size, neighbourhood threshold T, substitution matrix, masking, composition correction, gap costs and the E-value threshold together determine sensitivity and speed.
4. A homologue is missed if no seed passes the filter, even if its optimal local alignment would be significant: typically short, gap-rich alignments without a conserved stretch.
5. MMseqs2 also uses a k-mer prefilter (with similar, not only identical, k-mers) followed by ungapped and gapped alignment, but is designed for very large data sets with vectorisation and parallelisation.
6. Compare methods on the same frozen database and hardware, with a benchmark of known relationships, and compare sensitivity at the same error rate as well as running time and memory.

Common errors:

- Assuming that a heuristic computes an approximate score for every target; most targets never reach the alignment step.
- Treating a named program as having one fixed sensitivity.
- Comparing running times without the same database, hardware and sensitivity.

Connections: Section 1.4; benchmarking search methods (Section 1.11); indexing for read mapping (Section 5.4).

1.16.5 Control an iterative profile search

Explain how PSI-BLAST can detect remote homologues, and describe controls that limit profile drift during the iterations.

Extensions, solution outline and common errors

Extensions:

- Define what a position-specific scoring matrix (PSSM) represents.
- Explain the roles of sequence weighting and pseudocounts.
- Distinguish an inclusion threshold from a reporting threshold.
- When should the iterations be stopped?

Solution outline:

1. Start with an ordinary BLAST search, align the hits below the inclusion threshold to the query, and estimate position-specific log-odds scores.
2. Weight sequences to reduce the influence of densely sampled subfamilies, and add pseudocounts to regularise columns with few sequences.
3. Search again with the PSSM, and include only hits that pass the inclusion threshold.
4. Check the new hits of each round: do they cover the whole domain, do they have a compatible architecture and composition, and do they keep the conserved residues?
5. Mask low-complexity regions, use composition-based statistics, and keep the profile and the list of included sequences from each round.
6. Stop at convergence or when new hits become doubtful, and confirm the final set with an independent method, such as a curated profile HMM, structure comparison or a phylogenetic analysis.

Common errors:

- Including every displayed hit in the next round.
- Letting one shared domain justify a full-length multiple alignment.
- Assuming that more iterations always give better results.

Connections: Section 1.6; redundancy and sampling bias in databases (Section 2.2); paralogy and orthology (Section 3.4); annotation transfer (Section 10.4).

1.16.6 Explain profile hidden Markov models

What does a profile hidden Markov model represent that a PSSM does not, and how does this difference affect remote-homology detection and the interpretation of hits?

Extensions, solution outline and common errors

Extensions:

- Define match, insert and delete states and what each emits.
- Contrast Viterbi and forward scores.
- Distinguish per-domain from per-sequence significance in HMMER output.
- How are family-specific thresholds such as Pfam's gathering threshold set?

Solution outline:

1. A PSSM represents position-specific residue preferences; a profile HMM also represents position-specific probabilities of insertions and deletions.
2. Match states emit residues in consensus columns, insert states emit extra residues, and silent delete states skip consensus columns.
3. The Viterbi algorithm finds the single most probable state path; the forward algorithm sums over all paths, which is more sensitive when the alignment of a remote homologue is uncertain.
4. Separate the evidence for one domain from the summed evidence over a multidomain target; independent and conditional domain E-values answer different questions.
5. The model depends on a curated, representative seed alignment, sequence weighting, priors and the alignment mode (local or glocal).
6. Thresholds are set by curators so that no known false positives are included; domain membership and specific function must still be validated separately.

Common errors:

- Saying that delete states emit gap characters.
- Treating a significant domain match as proof of the function of the whole protein.
- Treating a profile as objective while ignoring the curation of its seed alignment.

Connections: Section 1.6; Pfam and InterPro (Section 1.9); structural families (Section 9.6); domain-based function prediction (Section 10.4).

1.16.7 Diagnose low complexity and domain rearrangement

The strongest database hits to a query align only to a low-complexity region rich in coiled coils. How would you distinguish a biologically informative relationship from an artefact of composition?

Extensions, solution outline and common errors

Extensions:

- Explain why biased composition changes the distribution of scores expected by chance.
- Contrast hard masking, soft masking and composition-based score correction.
- When can masking remove a real signal?
- How does the domain architecture change what is being tested for homology?

Solution outline:

1. Treat the result as uncertain, since biased composition can give high scores between unrelated segments when the background model does not fit.
2. Map low-complexity regions, repeats, coiled coils, transmembrane helices, disordered segments and candidate globular domains before searching again.
3. Apply masking (for example SEG) and composition-based statistics, and compare masked with unmasked results.
4. Search the candidate domains separately, and report the coverage of both query and target for every hit kept.
5. Look for supporting evidence outside the biased region: profiles, structure, conserved gene order, phylogeny or functional residues.
6. If the biased region remains the only evidence, report the relationship as unresolved.

Common errors:

- Declaring every low-complexity match an artefact.
- Masking without recording what was masked and how.
- Inferring full-length homology from a repeat shared by proteins with different architectures.

Connections: Section 1.4, Section 1.9; repeats in genomes (Section 6.7); disorder and structure prediction confidence (Section 9.8); interactions through coiled coils (Chapter 11).

1.16.8 Combine sequence, structure and embedding search

Sequence searches fail to find a plausible relative of a protein with a predicted structure. Compare profile, Foldseek and protein-language-model embedding searches, and describe how you would combine their evidence.

Extensions, solution outline and common errors

Extensions:

- Explain how Foldseek turns a three-dimensional structure into a searchable sequence.
- When may structural similarity reflect convergence rather than common ancestry?
- Define an embedding nearest neighbour, and explain why an embedding distance is not an evolutionary distance.
- How should a remote-homology benchmark be designed to avoid information leakage?

Solution outline:

1. Choose the next method from the available evidence: a family alignment supports a profile search, a reliable structure supports a structure search, and embedding search can propose candidates for sequences alone.
2. Foldseek describes each residue by a 3Di letter for its tertiary interactions with a neighbour, then uses MMseqs2-style prefiltering and alignment; check the aligned fold coverage, topology and confidence (pLDDT) of the model.
3. Physical constraints and convergent evolution can give similar folds without detectable common ancestry.
4. An embedding is a vector produced by a model and a pooling rule; its neighbours depend on the model, the distance measure and the reference database.

5. Check all candidates with residue-level alignment, profiles, structure, phylogeny and functional residues.
6. Benchmark with splits by superfamily or cluster rather than at random, with frozen databases and a check for overlap with the training data.

Common errors:

- Treating a predicted structure as an experimental observation, ignoring confidence and disorder.
- Inferring homology from a shared fold alone.
- Reporting performance on random splits as evidence of generalisation to remote families.

Connections: Section 1.4, Section 1.10, Section 1.11; predicted structures and their confidence (Section 9.8); foundation models (Section 12.6); function prediction (Section 10.4).

1.16.9 Interpret conservation in a non-coding region

A patient carries a de novo variant 40 kb upstream of a developmental gene. The base has a high phyloP score and lies in a phastCons element in the 241-mammal alignment. Explain what these scores measure, and what they do and do not tell you about the variant.

Extensions, solution outline and common errors

Extensions:

- Why can a mammalian genome alignment not be computed like a Needleman–Wunsch alignment of two genes?
- What are chains and nets, and how do they separate orthologous from paralogous alignments?
- How do phastCons, phyloP and GERP differ in what they report?
- What would a strongly negative phyloP score on the human branch mean?

Solution outline:

1. A whole-genome alignment must handle rearrangements, duplications and repeats: local matches are found (for example with LASTZ), chained into collinear blocks and filtered into nets, and combined into a multiple alignment with MULTIZ or a reference-free method such as Progressive Cactus.
2. Conservation is measured against a neutral model estimated from, for example, fourfold degenerate sites: phyloP tests each base for fewer (or more) substitutions than expected, phastCons gives the posterior probability that a base lies in a conserved element from a phylo-HMM, and GERP counts rejected substitutions.
3. A high score means that the base has been under purifying selection across mammals, so some function is likely and a change is more likely to matter.
4. It does not show which gene the element regulates, whether the variant changes its activity, or that the function is the same in humans; experimental evidence (reporter assays, chromatin data, segregation, phenotype match) is needed.
5. Check the alignment locally: missing species, paralogous alignment or an assembly gap can create apparent conservation.

Common errors:

- Treating conservation as proof of pathogenicity.
- Treating lack of conservation as proof that a region has no function (lineage-specific elements).
- Ignoring the depth of the alignment: a few primates cannot give base-level resolution.

Connections: Section 1.8; synteny and rearrangements (Section 6.5); how much of the genome is functional (Section 6.9); variant effect prediction (Section 10.6).

2 Databases

> **From the exam description**
>
> Knowledge of bioinformatics' most important databases and understanding their hierarchy and relationships.

> **This chapter is an overview, and will date**
>
> Databases change constantly: new resources appear, others merge or close, and sizes grow (or, as for UniProt in 2025, shrink by design). The numbers in this chapter are snapshots from October 2026 and are given only to show orders of magnitude. What matters for the exam is *what kind* of data each resource holds, where the data come from, how reliable they are, and how the resources relate to each other, not the current number of entries. For up-to-date descriptions, see the latest *Nucleic Acids Research* database issue (Rigden and Fernández 2025).

2.1 Overview

Bioinformatics depends on public databases, and knowing what each one contains, where its data come from and how much to trust it is a core skill. This chapter starts with the general principles: primary and secondary databases, curation and error propagation, identifiers, releases, licences and FAIR data, and how data and literature are retrieved (Section 2.2). It then goes through the main groups of resources:

- sequence databases: the nucleotide archives, UniProt, metagenomic sequences and protein families (Section 2.3);
- genome databases and browsers, annotation formats and coordinate conventions, and taxonomy databases (Section 2.4);
- experimental and predicted structures (Section 2.5);
- function and pathways: GO, KEGG, Reactome, ontologies and knowledge graphs, and how function prediction is evaluated (Section 2.6);
- interactions and networks (Section 2.7);
- expression and proteomics repositories (Section 2.8);
- medical, variant and cancer data, and controlled access to personal data (Section 2.9).

2.2 How biological databases work

2.2.1 Primary and secondary databases

It is useful to divide databases into two kinds.

- **Primary (archival) databases** store data as submitted by the people who produced it: nucleotide sequences, sequencing reads, structures, expression measurements, mass spectra. Their job is to preserve the record. The submitter, not the database, is responsible for the content, and errors are corrected only if the submitter updates the entry.
- **Secondary (derived or curated) databases** build on the primary data. They remove redundancy, add annotation, classify, integrate and interpret. Examples are RefSeq and UniProtKB/Swiss-Prot for sequences, Pfam and InterPro for families, SCOP and CATH for structures, and Reactome for pathways.

The secondary databases are themselves of two kinds. **Curated knowledgebases**, such as UniProtKB/Swiss-Prot, Reactome and the model organism databases, are built by experts who read the literature and record the evidence for each statement. **Derived resources** are computed from the archives with little or no manual review: clustered sets such as UniRef, automatic annotation such as UniProtKB/TrEMBL, precomputed family matches, and collections of predicted structures. Two pairs are often confused. GenBank, ENA and DDBJ form the primary archive of submitted nucleotide sequences, while RefSeq is a secondary, NCBI-maintained selection built from them. Swiss-Prot and TrEMBL are both part of UniProtKB, but only Swiss-Prot is manually reviewed (Section 2.3). The word "database" covers all of these; what matters is who produced the content and how it was checked.

Most major journals require that new sequences, structures and many other data types are deposited in the appropriate primary archive before a paper is published. This rule, which began with nucleotide sequences and the PDB in the 1980s and 1990s, is the reason why these archives are close to complete.

Databases are also organised along the **central dogma**. There are separate, cross-referenced resources for DNA (genomes and their annotation), RNA (transcripts) and protein (sequences, families, structures), and for what proteins do (functions, pathways, interactions). Knowing which *level* you need is the first step of any search.

2.2.2 Curation, automatic annotation and error propagation

Only a small fraction of database entries has been examined by a person. In 2026, UniProtKB/Swiss-Prot contains about half a million manually reviewed proteins, while the automatically annotated part, TrEMBL, contains more than a hundred million (Section 2.3). For most sequences, the annotation has been transferred automatically from a similar entry, which itself may have been annotated automatically.

This creates the risk of **error propagation**. A wrong function assigned to one protein is transferred to its homologues, then to theirs, and soon appears in thousands of entries that seem to support each other. Errors of this kind are common for multidomain proteins, where a hit to one domain is taken as evidence for the function of the whole protein (Section 1.9), and for paralogues with different functions. Good practice is therefore to check the **evidence** behind an annotation: whether it comes from an experiment on that protein, from curated literature, or from an automatic method. Evidence codes in the Gene Ontology (Section 2.6) and the "reviewed" status in UniProt make this possible.

Redundancy is another problem. Some proteins and genomes are sequenced thousands of times (for example, strains of *E. coli* or SARS-CoV-2), while most organisms are represented by a single genome or not at all. Search results and statistics can therefore be dominated by a few well-studied groups. Clustered and non-redundant databases, such as UniRef or the reference proteomes of UniProt, reduce this effect.

2.2.3 Records as evidence

A useful way to read any database entry is as a **versioned claim**: a statement about a biological entity, made by an identifiable source, on the basis of some evidence, in a particular release. Three layers are mixed on a typical gene or protein page. The **data** are the observations themselves (bases, reads, spectra, atomic coordinates). The **metadata** describe the context needed to interpret them: organism, sample, assay, instrument, reference assembly, processing. **Knowledge assertions** connect entities to interpretations, such as "protein P catalyses reaction R" or "variant V is pathogenic for condition C". An experimentally determined sequence, a predicted gene model, a curated function and an automatically transferred GO term can all appear side by side, in the same font.

Before trusting a statement, it should therefore be traced back to where it came from: who submitted or curated it, what it was computed from, and which release it belongs to. Several consequences follow.

- **Agreement is not independence.** If three databases call a protein a kinase, but two of them copied the annotation from the third, or all three transferred it from the same characterised homologue, there is one piece of evidence, not three. What counts is the number of independent experimental sources.
- **Absence is not evidence of absence.** A missing record means "not represented under this query in this release", not that the molecule, interaction or variant does not exist. What is in the databases is strongly biased towards model organisms, abundant and easily studied proteins, and well-funded diseases.
- **"Primary" is not a quality label.** Archived data can contain submitter errors, contamination or wrong sample labels, and a derived resource can be more reliable than the archive it is built from. The roles describe where the data come from, not how good they are. One resource can also play several roles: UniProtKB stores sequences translated from the archives, computes automatic annotation, curates the literature and integrates cross-references.
- **Test the evidence policy.** A simple robustness check is to repeat an analysis under two defensible policies, for example with reviewed UniProtKB entries only and with all entries, or with experimental GO evidence only and with all evidence codes. If the conclusion changes, the choice of evidence is part of the result.

2.2.4 Identifiers: accession numbers and names

Four kinds of identifier are easily confused:

- An **accession number** is a stable identifier for a record, unique *within one database*. It is kept when the record is updated, and the same gene has different accessions in GenBank, RefSeq, Ensembl and UniProt. Old accessions are kept as secondary accessions when records are merged.
- A **versioned accession** (`accession.version`) identifies one exact sequence. The version increases when the sequence changes, but not when only the annotation changes. For example, `NM_000518.5` is version 5 of the RefSeq mRNA for human β-globin. Ensembl uses a similar convention, in which the version of a gene or transcript identifier (such as `ENSG00000244734` for β-globin) increases when its structure changes. A result should always cite the accession *with* its version, because the sequence behind an accession without a version can change. UniProt keeps the version separately, as the *sequence version* of an entry.
- Other **database-specific identifiers and names** are meant for people rather than for stable reference. The UniProt entry name `HBB_HUMAN` can change, while the accession `P68871` does not. The **official gene symbol** (for human genes, assigned by the HUGO Gene Nomenclature Committee, HGNC, which also gives each gene a stable `HGNC:` identifier) is shared across databases, but symbols can change, and the same symbol may have been used for different genes in the past.

- **Persistent identifiers** such as the **DOI** (digital object identifier) are designed to resolve to an object for the long term, wherever it is hosted. DOIs identify papers, but also data sets (for example deposits in Zenodo) and PDB entries.

Cross-references connect the records for the same molecule in different databases. The β-globin gene, for example, is NCBI Gene `3043` and HGNC `HGNC:4827`, lies on chromosome 11 (`NC_000011.10`), has the mRNA `NM_000518.5` and the protein `NP_000509.1` in RefSeq, the UniProt entry `P68871` (HBB_HUMAN), the InterPro family `IPR000971` (Globin), and the structure `2HHB` in the PDB. Mapping between identifier types, with services such as UniProt ID mapping or Ensembl BioMart, is a routine but error-prone step. The mapping is rarely one-to-one: a gene has several transcripts and protein isoforms, a UniProt entry may merge several nucleotide records, and some identifiers are retired or split between releases. Unmapped or multiply mapped identifiers should be counted and reported, not silently dropped.

It helps to think of identifier mapping as a *join* between two tables rather than as replacing one string with another (Figure 2.1). Every mapping has a direction and a type (a transcript *translates to* a protein, a gene *has* a transcript), comes from a particular mapping release, and falls into one of four classes: one-to-one (still dependent on versions), one-to-many (isoforms, paralogues, split records), many-to-one (redundant records, aggregation), or unmapped (retired identifiers, the wrong namespace, a release mismatch). The rule used to resolve ambiguous cases should depend on the question: summing transcripts to genes may be right for gene-level expression, but it destroys the object of study in an analysis of isoform-specific function. Taking the first hit is never a rule.

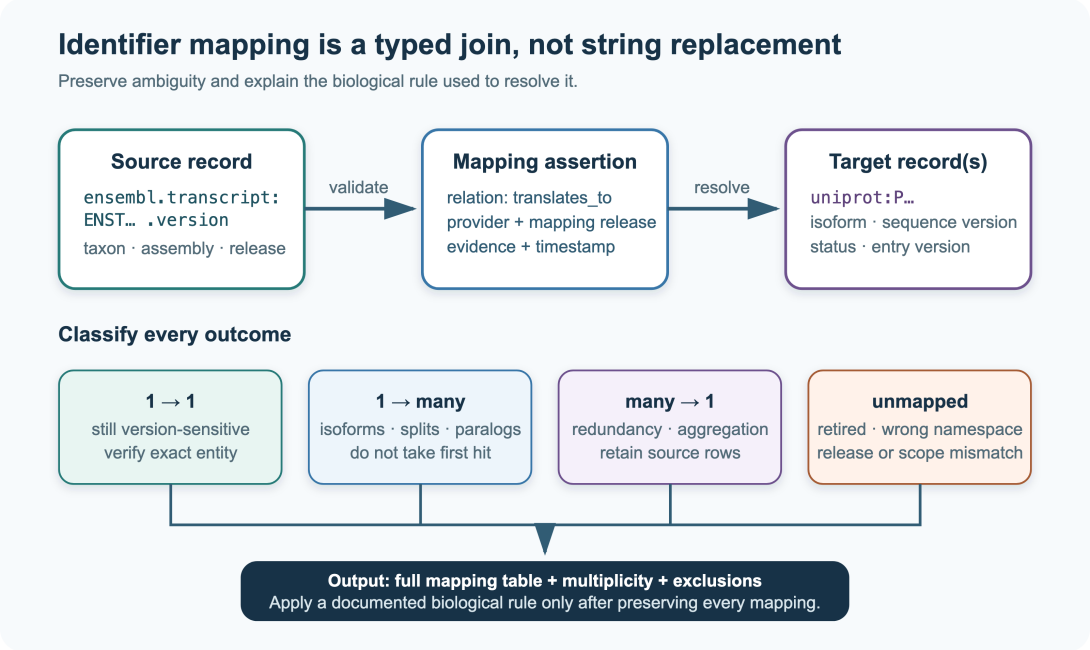


Figure 2.1: Identifier mapping as a typed join. A source record (here an Ensembl transcript with its version) is mapped through an explicit relation, with its provider and release, to one or more target records (here UniProt entries). Every outcome is classified, and the full mapping table, including multiple mappings and failures, is kept. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

Three further pitfalls are common:

- An accession only means something within its **namespace**: `P04637` is a UniProt accession, `GO:0006915` a GO term. **Compact identifiers** of the form `prefix:accession` (for example

`uniprot:P04637`) make the namespace explicit and can be resolved by services such as Identifiers.org (Wimalaratne et al. 2018).

- A **genomic coordinate is not an identifier**. `chr7:117,559,590` is ambiguous without the assembly (GRCh37 or GRCh38), the contig accession, whether positions are counted from 0 or 1, and how the alleles are written (Section 2.4). Converting coordinates between assemblies is a transformation that can fail, not a lookup.
- **Spreadsheets corrupt identifiers.** Opened in a spreadsheet program, gene symbols such as *SEPT2* or *MARCH1* were silently converted into dates, and long numeric identifiers into floating-point numbers. A survey found such errors in about a fifth of the papers whose supplementary spreadsheets contained gene lists (Ziemann et al. 2016); HGNC has since renamed the affected genes (to *SEPTIN2* and *MARCHF1*). Identifiers should be kept as text, and ideally not handled in spreadsheets at all.

2.2.5 Retrieval: Entrez, EBI Search and APIs

The large data centres provide integrated search systems over all their databases.

- **Entrez** is NCBI's search system. It links some 40 databases (nucleotide, protein, gene, genome, structure, taxonomy, PubMed and others) so that one can move from a paper to the sequences it describes, and from a gene to its protein, variants and structures (Sayers et al. 2024).
- **EBI Search** plays the same role for the resources at EMBL-EBI (ENA, UniProt, InterPro, PDBe, Ensembl, Expression Atlas and many others).
- Almost every major database also has a **programmatic interface (API)**, usually a REST service that returns data in JSON or a standard file format. Examples are NCBI's E-utilities, the UniProt and Ensembl REST APIs, and the RCSB PDB search and data APIs. Programmatic access is how large analyses are done, and it makes them easier to reproduce, because the exact query and database release can be recorded (Programming and reproducible research).

Many resources can also be downloaded in full. For large analyses this is often the best option, since it fixes the database version used.

A programmatic query should be treated like any other method. Record the literal query, the filters and the date; check that the number of records retrieved matches the number the server reports (results usually come in pages, and an interrupted download is easy to miss); keep the raw responses; and respect the service's usage rules. NCBI's E-utilities, for example, allow three requests per second without an API key and ten with one. Documented APIs and bulk downloads are preferable to extracting data from web pages, whose layout can change at any time.

2.2.6 Releases, versions and reproducibility

Databases are not fixed objects. UniProt, Pfam, Ensembl, GENCODE and most other resources appear in numbered or dated **releases**, and between releases entries are added, corrected, merged and deleted. The same query can therefore give different answers a year apart. A BLAST E-value depends on the size of the database searched (1), a GO enrichment depends on the annotations in that release, and the 2025 reduction of UniProtKB (Section 2.3) changed the content of every search against it. The database release, or the download date for resources without releases, is therefore part of every computational result, in the same way as the software version and the parameters. Record it in the methods, and keep or deposit the exact files used when the analysis must be reproducible (Programming and reproducible research).

2.2.7 Licences, FAIR data and access

Most core resources are free to use, but their **licences** differ. INSDC data carry no restrictions on use, PDB data are released under CC0, and UniProt under CC BY 4.0, which requires attribution. Some resources, such as KEGG (Section 2.6), OMIM and COSMIC (Section 2.9), are free to browse but require a licence for bulk download or commercial use. The licence matters when data are redistributed, for example as part of a derived database or a training set for a machine-learning model.

The **FAIR principles** state that data should be *findable*, *accessible*, *interoperable* and *reusable* (Wilkinson et al. 2016). In practice this means persistent identifiers, rich metadata, standard formats and vocabularies (such as the ontologies in Section 2.6) and a clear licence. FAIR does not mean open: data can be FAIR and still be under controlled access. Human genetic and health data are the main example. They are deposited in controlled-access archives such as dbGaP and EGA, where use is limited by the consent of the participants (Section 2.9).

2.2.8 The literature: PubMed, PMC and preprints

The scientific literature is itself a database. **PubMed**, run by NCBI, indexes the biomedical literature from MEDLINE and other sources: about 41 million records in October 2026. Records are indexed with Medical Subject Headings (MeSH), and searches can be combined with Boolean operators (`AND`, `OR`, `NOT`) and exact phrases in quotation marks. **PubMed Central** (PMC) is the free full-text archive, with about 12.7 million articles, roughly a third of PubMed. **Preprint servers**, above all bioRxiv for biology, make papers available before peer review. Preprints are now a normal part of the literature in bioinformatics, but they have not been reviewed and should be read with that in mind.

Every January, *Nucleic Acids Research* publishes a **database issue** with papers on new and updated databases (182 papers in 2026), and maintains a collection of more than 2,000 molecular biology databases (Rigden and Fernández 2025). It is the best place to find the current description of nearly every resource in this chapter.

2.2.9 Choosing a resource for a question

The rest of this chapter goes through the resources by data type. In practice the choice starts from the question, and each question has its own pitfalls:

Question	Where to start	What to check
What sequence was submitted?	INSDC: GenBank, ENA, DDBJ (Section 2.3)	Accession and version, sample metadata, contamination
What is the reference transcript or protein?	RefSeq, Ensembl (Section 2.4)	Assembly and annotation release, transcript version, MANE status
What is known about a protein?	UniProtKB	Reviewed or not, which isoform, evidence for each statement
Which domains or families does it contain?	InterPro and its member databases	Which method, match coordinates and thresholds, architecture
Is there an experimental structure?	PDB (Section 2.5)	Method, resolution and validation, construct, biological assembly
Is there a predicted structure?	AlphaFold DB and other model archives	That it is a prediction; model version; pLDDT and PAE

Question	Where to start	What to check
What does it do, in which pathway?	GO, Reactome, KEGG (Section 2.6)	Evidence codes, ontology release, whether pathway membership was transferred from another species
What does it interact with?	IntAct, BioGRID, STRING (Section 2.7)	Physical interaction or functional association; evidence type
Where and when is it expressed?	GEO, Expression Atlas, GTEx (Section 2.8)	Study design, sample metadata, raw versus processed data
Is a variant common, and has it been interpreted?	gnomAD, ClinVar (Section 2.9)	Ancestry and coverage; review status, condition and conflicts

The same molecule is represented differently in each of these resources (Figure 2.2). A submitted genomic sequence can give several transcripts; the canonical UniProt sequence need not be the construct that was crystallised; a pathway assignment may have been inferred from another species. The organism, accession and version, assembly, isoform or construct, the type of each link and the release must therefore be carried along at every step. The question is not just "where is this protein?" but "which representation of this protein supports which claim?"

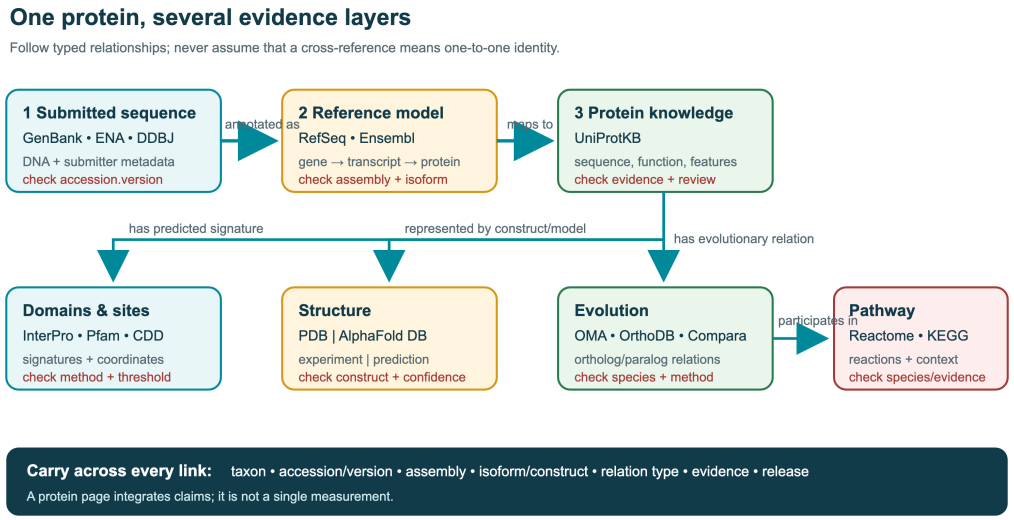


Figure 2.2: One protein seen through several evidence layers: the submitted sequence, the reference gene, transcript and protein models, the UniProtKB entry, and from there domains and sites, experimental or predicted structures, orthologues and pathways. The arrows are typed relationships, not statements that the records are identical. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

See also: further explanation in Section D.2.1; related Wikipedia articles in Section E.2.1.

Sources

Parts of this section are adapted from the Wikipedia articles Biological database, Accession number (bioinformatics), Entrez and PubMed (CC BY-SA 4.0, retrieved 2 October 2026). The structure and the β-globin example follow the database session of the KB8029 course book. Counts were

retrieved from the NCBI E-utilities on 2 October 2026. Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Biological Data, Databases, and Provenance" (CC BY 4.0), used with permission.

2.3 Sequence databases

2.3.1 The nucleotide archives: INSDC

All published nucleotide sequences are archived by the **International Nucleotide Sequence Database Collaboration (INSDC)** (Arita et al. 2020). It has three partners, which exchange their data every day so that the same records are available from all three:

- **GenBank** at NCBI in the USA, founded in 1982 at Los Alamos;
- the **European Nucleotide Archive (ENA)** at EMBL-EBI in the UK, whose predecessor, the EMBL Data Library, started in 1980;
- the **DNA Data Bank of Japan (DDBJ)**, founded in 1986.

INSDC is therefore one shared collection with three access points, not three independent databases: searching or counting all three gives the same records, not three times as many. Researchers submit to any of the three, and the record receives an accession number that is valid in all. The archives are not curated in the sense of being corrected by the database. An INSDC record is the submitter's sequence and annotation, and stays that way unless the submitter updates it. Data are free to use without restriction. The archives have grown with sequencing technology: GenBank release 271 (April 2026) contained more than 50 trillion bases in over six billion records, most of them from whole-genome shotgun and other bulk projects.

Raw sequencing data are kept separately, in the **Sequence Read Archive** (SRA) at NCBI and in ENA, which also mirror each other. A sequencing study is organised in a hierarchy of *study*, *sample*, *experiment* and *run* records, so that reads can be traced back to the biological samples they came from. Assembled genomes, transcriptome assemblies, and metadata about samples (BioSample) and projects (BioProject) are linked to the reads. Raw data are essential for reanalysis, but they are now so large that the archives store some of them in compressed or reduced forms.

2.3.2 Curated nucleotide and gene resources

Because the archives contain many redundant, partial and differently annotated records, curated resources have been built on top of them. **RefSeq** at NCBI, started in 2000, is such a secondary resource, not part of the INSDC archive. It provides a single, non-redundant, annotated record for each genomic sequence, transcript and protein of a selected organism (O'Leary et al. 2015). The accession prefix shows the type of molecule: `NC_` (complete chromosome), `NM_` (mRNA), `NR_` (non-coding RNA), `NP_` (protein), and `XM_`/`XP_` for transcripts and proteins predicted by the annotation pipeline but not yet supported by curated evidence. **NCBI Gene** collects information about each gene, and **Ensembl** provides the European counterpart for genomes and their annotation (Section 2.4).

2.3.3 Protein sequences: UniProt

Almost no protein sequences are determined directly any more. They are translated from genes predicted in sequenced genomes. **UniProt**, the Universal Protein Resource, is the central database for protein sequences and their functions (Bateman et al. 2024). It was formed in 2002 by combining

three older resources: **Swiss-Prot**, started by Amos Bairoch in 1986 in Geneva; its automatically annotated supplement **TrEMBL** (1996); and the **Protein Information Resource (PIR)**, the descendant of Margaret Dayhoff's *Atlas of Protein Sequence and Structure*. It is run by EMBL-EBI, the Swiss Institute of Bioinformatics and PIR. Its main parts are:

- **UniProtKB/Swiss-Prot**: manually reviewed entries. Curators read the literature, record function, catalytic activity, subcellular location, domains, post-translational modifications, variants and disease associations, and merge all sequences of the same protein from one species into one entry. About 576,000 entries (release 2026_03).
- **UniProtKB/TrEMBL**: unreviewed entries, annotated automatically by rule-based systems (UniRule, ARBA) that transfer annotation from reviewed entries to proteins of the same family. About 149 million entries (release 2026_03).
- **UniParc**: an archive in which every unique protein sequence from any source appears exactly once, under a stable identifier. It contains more than a billion sequences (2026).
- **UniRef**: clusters of UniProt and selected UniParc sequences at 100%, 90% and 50% identity (UniRef100, UniRef90, UniRef50) (Suzek et al. 2014). Searching UniRef instead of the full database removes most redundancy and is much faster. UniRef50 was, for example, used to train protein language models (Section 1.10).
- **Proteomes**: the set of proteins of each sequenced organism, with a selected **reference proteome** for each species or strain.

The difference in quality between the two sections of UniProtKB is large. A Swiss-Prot entry is a short review of what is experimentally known about a protein. A TrEMBL entry is a prediction: its sequence depends on an automatic gene model, and its annotation is inferred from homologues. The size of UniProtKB also depends on policy. To limit redundancy, UniProt decided in 2025 to keep only proteins from reference proteomes in UniProtKB, unless there is a specific reason to keep an entry, for example an experimental structure. This reduced UniProtKB from about 250 million to about 150 million entries during 2025–2026. The removed sequences remain available in UniParc.

2.3.4 Metagenomic sequences

A growing share of all known protein sequences comes not from cultured organisms but from **metagenomes**: DNA sequenced directly from environmental or host-associated samples. Most of these sequences are not in UniProtKB. **MGnify** at EMBL-EBI assembles and annotates metagenomic data and provides a protein database with billions of sequences (Richardson et al. 2022). Clustered collections of metagenomic and genomic proteins, such as the Big Fantastic Database (BFD) and the databases built with Linclust (Section 1.4), are used to build the deep multiple alignments needed for structure prediction (Chapter 9). Metagenomic proteins have also been predicted structurally on a massive scale (Section 2.5).

2.3.5 Protein families and domains

Family and domain databases classify sequences by homology. They are described in detail in Section 1.9 and are listed here only for completeness:

- **Pfam**: families and domains defined by curated seed alignments and profile HMMs, now including deep-learning annotations (Pfam-N) (Mistry et al. 2020; Paysan-Lafosse et al. 2024).
- **InterPro**: integrates Pfam and about a dozen other signature databases (PROSITE, SMART, PANTHER, CDD, NCBIfam, CATH-Gene3D, SUPERFAMILY and others) into a single classification. Its matches are precomputed for all of UniProtKB (Blum et al. 2024).

- **SCOP, CATH, ECOD and TED**: classifications of domains of known or predicted three-dimensional structure (Section 2.5).
- **Rfam**: the corresponding database for families of non-coding RNAs (Section 1.7).

See also: further explanation in Section D.2.2; related Wikipedia articles in Section E.2.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles International Nucleotide Sequence Database Collaboration, GenBank, European Nucleotide Archive, Sequence Read Archive, RefSeq and UniProt (CC BY-SA 4.0, retrieved 2 October 2026). UniProt counts were retrieved from the UniProt REST API (release 2026_03) on 2 October 2026.

2.4 Genome databases and browsers

A genome sequence becomes useful when it is annotated: genes, transcripts, regulatory elements, variants and comparisons with other species have to be placed on the sequence and kept up to date as assemblies improve. **Genome browsers** present this information graphically, as tracks along a chromosome, and are among the most used tools in biology.

- **Ensembl** was started in 1999 by EMBL-EBI and the Sanger Institute to annotate the human genome automatically as it was being sequenced (Dyer et al. 2024). It now covers hundreds of vertebrate genomes, with gene models built by an automatic pipeline from aligned transcripts and proteins, combined with manual annotation (GENCODE) for human and mouse. It also provides comparative genomics (orthologues, paralogues and whole-genome alignments), variation data and the Variant Effect Predictor. **Ensembl Genomes** extends the system to bacteria, fungi, plants, protists and invertebrates. Ensembl is moving to a new browser and data model designed for many more genomes.
- The **UCSC Genome Browser**, from the University of California, Santa Cruz, displayed the first working draft of the human genome in 2000 (Perez et al. 2024). It is known for its large collection of annotation tracks contributed by many projects (conservation, regulatory elements, expression, variation, repeats), its fast display, and tools such as BLAT (Section 1.4) and the Table Browser. Users can upload their own data as custom tracks.
- **NCBI** provides the Genome, Gene and RefSeq resources and its own Genome Data Viewer, together with the annotation pipelines that produce the RefSeq gene models for eukaryotes and prokaryotes (Sayers et al. 2024).

Ensembl, UCSC and NCBI annotate the human genome independently. Their gene sets largely agree, and the **MANE** project (Matched Annotation from NCBI and EMBL-EBI) now defines one agreed reference transcript per human protein-coding gene, which is important for reporting clinical variants consistently. The coordinates of features depend on the **genome assembly** (for example GRCh37/hg19 or GRCh38/hg38 for human), so the assembly version must always be stated. Converting coordinates between assemblies (“liftover”) is a common source of errors. The first complete telomere-to-telomere human genome (T2T-CHM13) and the human pangenome reference are gradually being added to these resources (Chapter 5).

2.4.1 Annotation formats and coordinates

Features on a genome, such as genes, exons, peaks or variants, are exchanged as tables of coordinates. Two families of formats dominate, and they count positions differently.

- **BED** (Browser Extensible Data), from UCSC, has three required columns (chromosome, start, end) and up to nine optional ones (name, score, strand, the coding region, colour, and the sizes and starts of blocks such as exons). It is compact and is the input of tools such as BEDTools (Quinlan and Hall 2010), which intersect, merge and compare sets of intervals.
- **GFF3** (General Feature Format) has nine columns: sequence, source, feature type, start, end, score, strand, phase (the reading frame of a CDS) and free-form attributes. Feature types are terms from the Sequence Ontology, and the `ID` and `Parent` attributes link exons and CDSs to transcripts and transcripts to genes, so that a whole gene model can be described. **GTF** is an older dialect (GFF version 2) with mandatory `gene_id` and `transcript_id` attributes, still used by Ensembl and GENCODE and by most RNA-seq quantification tools (Chapter 7).

The two families also use different **coordinate conventions**. BED is **0-based, half-open**: the first base of a chromosome is position 0, and the interval $[s, e)$ includes the start but not the end, so its length is $e - s$. GFF3, GTF, VCF and SAM are **1-based, closed**: the first base is 1, and $[s, e]$ includes both ends, so the length is $e - s + 1$. The same single base, the 100th of chromosome 1, is therefore written `chr1 99 100` in BED and as start = end = 100 in GFF. Converting between the two means subtracting 1 from the start and leaving the end unchanged. The UCSC browser displays 1-based coordinates but stores 0-based ones, and BAM files store 0-based positions that are displayed as 1-based SAM. Forgetting a conversion gives **off-by-one errors**: features shifted by one base, zero-length or one-base-too-long intervals, and variants assigned to the wrong base or the wrong codon. Such errors are silent: the files still parse, and the results are almost right. A related source of silent failure is chromosome naming (`chr1` at UCSC, `1` at Ensembl), which can make two files that describe the same genome share no features at all.

Coordinates also only mean something together with the **genome build**, the assembly version they refer to (Section 2.2). Human GRCh37 is the same assembly as UCSC's hg19, and GRCh38 the same as hg38; the positions of most features differ between them, sometimes by megabases. **Liftover** converts coordinates from one build to another using a whole-genome alignment of the two assemblies, stored as chain files (Section 1.8), with UCSC liftOver or CrossMap (Zhao et al. 2013). Liftover is a transformation, not an identity: some regions cannot be converted because they were changed, added or removed in the new assembly, others are split or map to several places, and alleles may differ when the reference base has changed. For important analyses it is better to map the original reads to the new assembly than to lift over the results.

2.4.2 Organism-specific databases

For each major model organism there is a dedicated **model organism database**, curated by experts who follow the literature on that organism. Examples are SGD (budding yeast), PomBase (fission yeast), FlyBase (*Drosophila*), WormBase (*C. elegans*), ZFIN (zebrafish), MGI (mouse), RGD (rat), TAIR (*Arabidopsis*) and EcoCyc (*E. coli*). They hold gene nomenclature, mutant phenotypes, genetic and physical interactions, and detailed functional annotation that the general databases often import. Several of them are federated in the Alliance of Genome Resources. For prokaryotes, collections of genomes and taxonomies such as NCBI's prokaryotic RefSeq and the Genome Taxonomy Database (GTDB), which classifies bacteria and archaea by genome phylogeny, are the most important resources (Chapter 6).

2.4.3 Taxonomy databases

Almost every database record carries the organism it came from, and the taxonomy is the key that links them. **NCBI Taxonomy** (Schoch et al. 2020) holds the names and classification of all organisms with sequences in INSDC (Section 2.3), each identified by a numerical **taxid** (9606 for *Homo sapiens*, 562 for *Escherichia coli*). The taxid, not the name, is what UniProt, RefSeq, Ensembl, the PDB and many other resources use to say which organism an entry belongs to, and it is how searches can be

restricted to a clade (for example, BLAST against Metazoa only). Names change and can be ambiguous; taxids are stable, although they can be merged or deleted when the classification is revised. The NCBI database explicitly states that it is not an authoritative source for nomenclature or classification: it is a curated working classification that follows the literature and taxonomic experts, and it contains many provisional names for unclassified and environmental organisms.

For bacteria and archaea, the **Genome Taxonomy Database** (GTDB) (Parks et al. 2018, 2021) offers an alternative: a classification derived entirely from a phylogeny of conserved single-copy proteins of genomes, including metagenome-assembled genomes, with ranks normalised by evolutionary divergence. It reclassified many groups and renamed many taxa, so the same genome can have different names in NCBI and GTDB. The principles behind the two approaches are discussed in Section 4.5.

See also: further explanation in Section D.2.3; related Wikipedia articles in Section E.2.3.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Ensembl genome database project, UCSC Genome Browser and Model organism database (CC BY-SA 4.0, retrieved 2 October 2026). The subsections on annotation formats and taxonomy databases were written for this book from the primary literature cited in the text.

2.5 Structure databases

Structure resources fall into two groups that should not be confused: the PDB, an archive of structures determined by experiment, and collections of computationally predicted models, such as the AlphaFold database.

2.5.1 The Protein Data Bank: experimental structures

The **Protein Data Bank** (PDB), founded in 1971 at Brookhaven National Laboratory with seven structures, is the single global archive of experimentally determined structures of proteins, nucleic acids and their complexes (Berman et al. 2000; Burley et al. 2018). Since 2003 it has been managed by the **worldwide PDB (wwPDB)**: RCSB PDB in the USA, PDBe in Europe, PDBj in Japan, the Biological Magnetic Resonance Data Bank (BMRB) and, since 2021, the Electron Microscopy Data Bank (EMDB). Depositions are processed and validated by curators, and the data are freely available under a CC0 licence. In October 2026 the PDB contained about 260,000 experimental structures.

Each entry contains the atomic coordinates, the experimental data (such as structure factors for X-ray crystallography or restraints for NMR), information about the sample and the method, and a **validation report** that compares the model with the data and with expected geometry. Quality varies a lot between entries. For crystal structures, the resolution and the R and R-free values are the first things to check (Brünger 1992). Note also that the deposited *asymmetric unit* is not necessarily the biologically relevant assembly, which is given separately. The standard file format is now PDBx/mmCIF, which has replaced the old fixed-column PDB format because that format cannot represent very large structures. **Integrative** (or hybrid) structures, which combine data from several experimental methods with computational modelling, use an extension of PDBx/mmCIF and are archived through PDB-IHM (formerly PDB-Dev), now part of the wwPDB. They are experimentally based, but their precision is often lower than that of a crystal or cryo-EM structure.

EMDB archives the three-dimensional density maps from cryo-electron microscopy and tomography (Turner et al. 2023). The atomic models built into the maps are deposited in the PDB. With the rapid

growth of cryo-EM, an increasing share of new structures, especially of large complexes and membrane proteins, comes from EMDB/PDB pairs.

Structures are classified into domains, folds and superfamilies by **SCOP**, **CATH** and **ECOD**, and the domains of predicted structures by **TED** (Section 1.9).

2.5.2 Predicted structures: model databases

Since 2021, predicted structures have become a resource in their own right, much larger than the PDB. These are prediction resources, not archives of experimental data, and purely computational models are not accepted in the PDB. They are models, not measurements, and must always be used together with their confidence estimates (pLDDT and predicted aligned error, PAE; Chapter 9).

- The **AlphaFold Protein Structure Database** (AFDB), built by DeepMind and EMBL-EBI, was released in July 2021 with the proteomes of humans and 20 model organisms, and expanded in 2022 to more than 200 million proteins, almost all of UniProt (Varadi et al. 2021, 2023). In 2025 it was updated to a recent UniProt release, with a redesigned interface, isoforms and domain annotations (Bertoni et al. 2025). In 2026 it started to include predicted **complexes**: about 80,000 high-confidence heterodimers and 1.7 million homodimers, mainly from about 20 well-studied species and priority pathogens, with millions of lower-confidence predictions available for download.
- The **ESM Metagenomic Atlas** contains more than 600 million structures of metagenomic proteins, predicted with ESMFold (Lin et al. 2023). Most of these proteins are from organisms that have never been cultured.
- AFDB does not cover viral proteins. The **Big Fantastic Virus Database** (BFVD), from Martin Steinegger's group, fills this gap with structures of viral proteins predicted with ColabFold (Kim et al. 2024). Its first version contained about 350,000 structures of representative sequences. Many of them have no structural similarity to anything in the PDB or AFDB, and later versions cover almost all viral proteins in UniProt.
- Other collections of models (ModelArchive, SWISS-MODEL Repository) are linked to the PDB and AFDB through the **3D-Beacons** network, so that experimental and predicted models of a protein can be found in one place.

These collections are only searchable at this scale because of fast structure-search tools such as Foldseek, which also made it possible to cluster the whole AFDB and discover new folds (van Kempen et al. 2023; Barrio-Hernandez et al. 2023) (Section 1.4).

See also: further explanation in Section D.2.4; related Wikipedia articles in Section E.2.4.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Protein Data Bank, EM Data Bank and AlphaFold (CC BY-SA 4.0, retrieved 2 October 2026). The PDB count was retrieved from the RCSB Search API on 2 October 2026; the 2026 AFDB complexes are described in an EMBL announcement.

2.6 Function and pathways

2.6.1 The Gene Ontology

The **Gene Ontology** (GO), started in 1998 by the model organism databases for fly, yeast and mouse, provides a shared, controlled vocabulary for describing what genes and proteins do in any organism

(Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023). It has three independent parts:

- **Molecular function**: the activity of the gene product at the molecular level, such as "protein kinase activity" or "ATP binding";
- **Biological process**: the larger biological programme that the activity contributes to, such as "DNA repair" or "cell division";
- **Cellular component**: where the gene product is located or acts, such as "mitochondrial inner membrane" or "ribosome".

Terms are organised as a **directed acyclic graph** rather than a strict tree: a term can have several parents, linked by relations such as *is_a* and *part_of*. An annotation to a specific term implies all of its ancestors (the "true path rule"). This structure is what makes GO useful for computation, for example for **enrichment analysis**, where one asks whether a list of genes is annotated to some term more often than expected by chance (Section 7.4.1, Chapter 7).

Every annotation of a gene to a GO term carries an **evidence code**. Annotations based on experiments (for example IDA, *inferred from direct assay*) are distinguished from those inferred from phylogeny or sequence similarity (IBA, ISS) and from fully automatic electronic annotations (IEA), which are the large majority. As discussed in Section 2.2, the evidence code tells you how much to trust an annotation, and whether it is independent evidence or a copy of another annotation. Other classification systems used alongside GO are the **Enzyme Commission (EC) numbers** for enzymatic reactions and the reaction database Rhea.

2.6.2 Pathway databases

- **KEGG** (Kyoto Encyclopedia of Genes and Genomes) was started in 1995 by Minoru Kanehisa at Kyoto University (Kanehisa et al. 2024). Its best-known part is a set of manually drawn **pathway maps** for metabolism, signalling, cellular processes and diseases. Genes are linked to the maps through **KEGG Orthology** (KO) groups, so that the pathways can be reconstructed automatically for any sequenced genome. KEGG also contains modules, compounds, reactions, drugs and diseases. The website is free for academic use, but bulk downloads require a licence, which is why some open tools use other resources.
- **Reactome** is an open, expert-curated database of human pathways and reactions, developed by OICR, EMBL-EBI, NYU and OHSU (Milacic et al. 2023). Pathways are built from individual reactions with defined inputs, outputs, catalysts and regulators, each supported by literature references. Pathways for other species are inferred by orthology. Reactome is widely used for pathway enrichment analysis.
- Other important resources are MetaCyc and BioCyc for metabolic pathways across organisms, WikiPathways for community-curated pathways, and BioModels for quantitative, computational models of pathways.

2.6.3 Ontologies and knowledge graphs

GO is one of many biomedical **ontologies**, controlled vocabularies whose terms are linked by defined relations. Others describe chemical entities (ChEBI), sequence features (Sequence Ontology), anatomy (Uberon), cell types (Cell Ontology), human phenotypes (Human Phenotype Ontology) and diseases (Mondo). Many are developed together in the OBO Foundry, so that they use compatible identifiers and relations. When databases annotate their records with the same ontology terms, the records can be combined across resources.

A **knowledge graph** takes this integration further. Genes, proteins, GO terms, pathways, compounds, diseases and phenotypes become nodes, and relations from many databases (a protein *participates in* a

Reactome reaction, a gene *is associated with* a disease, a drug *targets* a protein) become typed edges. Several resources, UniProt among them, also distribute their data in this form (as RDF, queried with SPARQL). Knowledge graphs are used for queries across databases, for prioritising disease genes and for drug repurposing (Himmelstein et al. 2017), often with graph machine-learning methods (Chapter 12). They inherit the biases of their sources: well-studied genes have many more edges, and a missing edge usually means "not studied" rather than "no relation".

2.6.4 Evaluating function prediction: CAFA

Because most annotations are predictions, it matters how good the prediction methods are. The **Critical Assessment of Functional Annotation** (CAFA) evaluates them prospectively. Participants predict GO terms for thousands of proteins without experimental annotation, and the predictions are scored against the experimental annotations that accumulate in the following months (Radivojac et al. 2013; Zhou et al. 2019). The fifth round (CAFA5) was run as a Kaggle competition with a much larger number of participants. It showed clear improvements, attributed to protein language models, predicted structures and the growth of experimental annotations (De Paolis Kaluza et al. 2026). CAFA also highlights a general problem of the field: annotations are incomplete, so a "false" prediction may simply be one that has not yet been tested (Chapter 10).

See also: further explanation in Section D.2.5; related Wikipedia articles in Section E.2.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Gene Ontology, KEGG and Reactome (CC BY-SA 4.0, retrieved 2 October 2026).

2.7 Interaction databases

Interaction databases differ in *what kind* of relationship they record, and this matters for how they can be used (Chapter 11).

- **Curated physical interactions. IntAct** at EMBL-EBI curates molecular interactions from the literature in fine detail: the experimental method, the participants, their roles and any binding regions (Toro et al. 2021). It is the main member of the **IMEx** consortium, which shares a common curation standard. **BioGRID**, started in 2003, curates protein–protein and genetic interactions, chemical interactions and post-translational modifications for the major model organisms and human, and includes many large screens (Oughtred et al. 2020). The **Complex Portal** describes curated stable complexes with their composition and stoichiometry, for example the haemoglobin tetramer of two α and two β chains (Meldal et al. 2021).
- **Functional association networks. STRING** combines experimental interactions with evidence from genomic context (gene neighbourhood, fusion, co-occurrence), co-expression, curated pathways and text mining, and transfers associations between organisms by orthology. Each link has a confidence score, and the network covers thousands of organisms (Szklarczyk et al. 2022, 2024). The 2025 version also distinguishes directed, regulatory links. **FunCoup**, developed by Erik Sonnhammer's group in Stockholm, builds networks with a naive Bayesian framework that integrates many types of evidence, and recent versions include directed gene-regulatory links for human and many other species (Buzzao, Persson, et al. 2024; Guala et al. 2019). A link in these networks means that two proteins probably *work together*. It does not necessarily mean that they bind each other.

- **Maps of complexes from proteomics. hu.MAP** integrates thousands of affinity-purification and co-fractionation mass-spectrometry experiments with machine learning to predict human protein complexes. Version 3.0 combined more than 25,000 experiments and identified more than 15,000 complexes covering about 70% of the human proteome (Fischer et al. 2025). The complexes are also imported into the Complex Portal.
- **Predicted structures of complexes.** Large sets of AlphaFold-predicted complexes are now available, including in the AlphaFold database (Section 2.5). They add a structural model and a confidence to a predicted interaction.

When using interaction data, ask what type of evidence a link is based on. High-throughput screens and text mining have different error rates, and interactions between very well-studied proteins are over-represented simply because more experiments have been done on them.

See also: further explanation in Section D.2.6; related Wikipedia articles in Section E.2.6.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles STRING and BioGRID (CC BY-SA 4.0, retrieved 2 October 2026).

2.8 Expression and proteomics repositories

Functional genomics experiments are archived in their own primary repositories, again usually as a condition of publication.

- **GEO** (Gene Expression Omnibus), at NCBI since 2000, archives microarray, RNA-seq, single-cell and epigenomics data sets, with processed data and descriptions of the samples, while the raw reads are stored in the SRA (Clough et al. 2023).
- **ArrayExpress**, at EMBL-EBI since 2002, is the European counterpart. It is now part of **BioStudies**, a general archive for data that belong to one study, with the raw reads in ENA.
- Curated, reanalysed views of these data are provided by, for example, the **Expression Atlas** and **Single Cell Expression Atlas** at EMBL-EBI, **GTEx** (gene expression across about 50 human tissues from around a thousand donors), and the **Human Protein Atlas**, a Swedish project started in 2003 that maps human proteins in tissues, cells and organs with antibodies, transcriptomics and other methods (Uhlén et al. 2015).
- **PRIDE** at EMBL-EBI is the largest repository of mass-spectrometry proteomics data (Perez-Riverol et al. 2024). It is part of the **ProteomeXchange** consortium, which coordinates submission and data sharing between proteomics repositories (Chapter 8).

The value of these archives depends heavily on the **metadata**: which samples were compared, how they were treated and how the data were processed. Reanalysing public data is a powerful approach, but incomplete or inconsistent sample descriptions are a common obstacle. Standards such as MIAME (for microarrays) and MINSEQE (for sequencing) define the minimum information that should be provided.

See also: further explanation in Section D.2.7; related Wikipedia articles in Section E.2.7.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Gene Expression Omnibus and Human Protein Atlas (CC BY-SA 4.0, retrieved 2 October 2026).

2.9 Medical and variant data

Human genetic and clinical data raise two issues that the databases above do not have: the interpretation of variants for health, and the protection of personal data. The resources therefore fall into two groups: open databases of variants and their interpretation, and controlled-access archives of individual-level data.

2.9.1 Variants and their interpretation

- **dbSNP** at NCBI, started in 1998, catalogues short genetic variants (single-nucleotide variants, small insertions and deletions) and gives each a stable identifier, the `rs` number. The **European Variation Archive** (EVA) plays a similar role at EMBL-EBI for variants from all species.
- **gnomAD** (Genome Aggregation Database) aggregates exome and genome sequences from large studies to provide allele frequencies in different populations. Version 4 contains data from more than 800,000 individuals (Chen et al. 2023). Its main use is to tell whether a variant is rare or common, and which genes tolerate loss-of-function variants (*constraint*), which is essential when judging whether a variant could cause a rare disease. The **1000 Genomes Project** produced an earlier, openly accessible reference of human genetic variation.
- **ClinVar** at NCBI collects interpretations of variants submitted by clinical laboratories and researchers, classified on a scale from *benign* to *pathogenic*, together with the supporting evidence (Landrum et al. 2017). Most clinical laboratories classify variants with the ACMG/AMP guidelines, which combine population frequency, segregation, functional data and computational predictions (Richards et al. 2015). A ClinVar classification is therefore a submitter's interpretation, not an experimentally established fact. Conflicting interpretations between submitters are common and are shown explicitly, and each record has a **review status** (zero to four stars), from a single submitter without stated criteria to an expert panel or practice guideline. Classifications also change over time, often from *uncertain significance* to benign or pathogenic. When ClinVar is used as a ground truth, for example to train or evaluate variant effect predictors, the review status should be filtered, the release recorded, and circularity with computational evidence kept in mind (Section 10.6, 4).
- **OMIM** (Online Mendelian Inheritance in Man) is the classic curated catalogue of human genes and genetic disorders, going back to Victor McKusick's *Mendelian Inheritance in Man* from the 1960s (Amberger et al. 2018).
- The **GWAS Catalog** (NHGRI and EMBL-EBI) collects the associations between variants and traits reported in genome-wide association studies (Sollis et al. 2022).
- Computational predictions of variant effects, for example AlphaMissense predictions for all possible human missense variants (Cheng et al. 2023), are integrated into several of these resources (Chapter 10).

These resources answer different questions, and mixing them up is a common error. An `rs` number in dbSNP shows that a variant has been *observed* before, not that it is benign. gnomAD gives its *frequency* in the sampled populations; it is not a collection of healthy controls (its cohorts include people with common diseases), and a variant's absence from gnomAD is weak evidence if coverage at that position is poor or the relevant ancestry is under-represented. ClinVar records *interpretations* for a stated condition and mode of inheritance, which should be read together with their review status and dates rather than counted as votes. All three depend on the reference assembly and on the transcript used to describe the variant: the same genomic change can have different consequences in different transcripts, and GRCh37 and GRCh38 coordinates must never be mixed.

2.9.2 Cancer genomics

- **The Cancer Genome Atlas** (TCGA), a project of the US National Cancer Institute and NHGRI from 2006, characterised more than 11,000 tumours from 33 cancer types with genome, exome, transcriptome, methylation and protein data (Weinstein et al. 2013). Its data, and those of later projects, are distributed through the NCI **Genomic Data Commons** (GDC). The International Cancer Genome Consortium (ICGC) coordinated similar projects worldwide.
- **COSMIC** (Catalogue of Somatic Mutations in Cancer), at the Sanger Institute since 2004, is an expert-curated database of somatic mutations in cancer, including the Cancer Gene Census of genes with a documented role in cancer (Sondka et al. 2023).

2.9.3 Controlled access to personal data

Sequences and clinical data from individual people can identify them, so they are not openly published. They are deposited in **controlled-access archives**, such as dbGaP at NCBI and the **European Genome-phenome Archive** (EGA) at EMBL-EBI and CRG in Barcelona. Researchers must apply to a data access committee and agree to conditions on how the data are used and stored. The permitted uses follow from the consent the participants gave: a data set may be restricted to research on a particular disease, to non-commercial use, or to approved projects only. These restrictions travel with the data and also apply to any derived data set. In Europe, the use of such data is regulated by the General Data Protection Regulation (GDPR); in Sweden, research on personal data also requires ethical approval (Programming and reproducible research). Large biobanks, such as the UK Biobank with genetic and health data from about half a million participants, are accessed in the same way, increasingly through secure analysis environments where the data never leave the provider.

Controlled access is compatible with the FAIR principles (Section 2.2): the existence and description of a data set can be open even when the data themselves are not. The **CARE principles** (*collective benefit*, *authority to control*, *responsibility* and *ethics*) complement FAIR for Indigenous data governance (Carroll et al. 2020). They stress that the peoples and communities from which data come should have a say in how the data are used and should benefit from that use.

This area changes particularly quickly, and many national resources (for example for Swedish health data) are not covered here. For a current picture, start from the NAR database issue (Rigden and Fernández 2025) and the websites of the resources themselves.

See also: further explanation in Section D.2.8; related Wikipedia articles in Section E.2.8.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles dbSNP, ClinVar, Online Mendelian Inheritance in Man, GWAS catalog, The Cancer Genome Atlas, COSMIC cancer database and European Genome-phenome Archive (CC BY-SA 4.0, retrieved 2 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Biological Data, Databases, and Provenance” (CC BY 4.0), used with permission.

2.10 Key concepts

A checklist for revision.

- **Principles:** primary archives vs. curated knowledgebases vs. derived resources (GenBank vs. RefSeq, Swiss-Prot vs. TrEMBL); the DNA–RNA–protein organisation; curation vs. automatic annotation; error propagation and evidence codes; redundancy and non-redundant sets; records as versioned claims (data, metadata, knowledge assertions); why agreeing databases are not independent evidence; absence of a record is not absence; testing robustness to the evidence policy
- **Identifiers and retrieval:** stable accessions vs. versioned accessions vs. names and symbols; persistent identifiers (DOI); cross-references and identifier mapping as a typed join (one-to-one, one-to-many, many-to-one, unmapped); namespaces and compact identifiers; why a genomic coordinate is not an identifier; spreadsheet corruption of gene symbols; Entrez, EBI Search and REST APIs; recording queries and checking that retrieval is complete; choosing a resource from the question; PubMed, MeSH, PMC and preprints; the NAR database issue
- **Releases and reuse:** the database release as part of every result; licences (CC0, CC BY, restricted bulk download); the FAIR principles; FAIR is not the same as open
- **Sequences:** INSDC (GenBank, ENA, DDBJ): one shared collection, exchanged daily; SRA and the study–sample–experiment–run hierarchy; RefSeq and its accession prefixes; UniProtKB/Swiss-Prot vs. TrEMBL, UniParc, UniRef and reference proteomes; metagenomic sequence collections (MGnify); family databases (Pfam, InterPro)
- **Genomes:** Ensembl, UCSC and NCBI; gene models and MANE; BED vs. GFF3/GTF; 0-based half-open vs. 1-based closed coordinates and off-by-one errors; chromosome naming; genome builds (GRCh37/hg19, GRCh38/hg38) and liftover with chain files, and why it can fail; model organism databases
- **Taxonomy:** NCBI Taxonomy and taxids as the organism key across databases (and not an authoritative nomenclature); GTDB's genome-based, rank-normalised classification of bacteria and archaea; why the same genome can have different names
- **Structures:** the PDB and wwPDB as the archive of *experimental* structures (integrative structures in PDB-IHM); what an entry contains, validation, resolution and R-free, asymmetric unit vs. biological assembly; mmCIF; EMDB; *predicted* models in AlphaFold DB (including complexes), ESM Atlas and BFVD; confidence measures of predicted models
- **Function and pathways:** the three GO aspects, the GO graph and the true path rule, evidence codes; EC numbers; KEGG (pathway maps, KO) and Reactome (reactions); other ontologies (OBO Foundry) and knowledge graphs; CAFA and why function prediction is hard to evaluate
- **Interactions:** curated physical interactions (IntAct/IMEx, BioGRID, Complex Portal) vs. functional association networks (STRING, FunCoup) vs. complex maps (hu.MAP); what a link means
- **Expression and proteomics:** GEO, ArrayExpress/BioStudies, Expression Atlas, GTEx, Human Protein Atlas; PRIDE and ProteomeXchange; the importance of metadata
- **Medical data:** dbSNP, gnomAD, ClinVar (submitted interpretations, review status, conflicts); observation vs. frequency vs. interpretation, and why gnomAD is not a set of healthy controls, OMIM, GWAS Catalog; TCGA/GDC and COSMIC; controlled access (dbGaP, EGA) and consent-based use restrictions, GDPR and ethical approval; the CARE principles

2.11 Recommended reading

These papers are recommended for reading in full:

- Bateman et al. (2024) — UniProt today; what Swiss-Prot, TrEMBL, UniParc and UniRef are, and how annotation is done.
- Blum et al. (2024) — InterPro, the integrated classification of protein families, domains and sites.
- Burley et al. (2018) — the Protein Data Bank as a single global archive, and how structures are deposited and validated.

- Bertoni et al. (2025) — the AlphaFold Protein Structure Database and how to use predicted structures.
- Szklarczyk et al. (2024) — STRING; how functional association networks are built and scored.
- The Gene Ontology Consortium et al. (2023) — the Gene Ontology knowledgebase; terms, annotations and evidence.

Titles, abstracts, links and the full texts are in Section A.2; a longer list of further papers is in Section B.2.

2.12 Background reading (open access)

- **Biological database (Wikipedia)** — Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis.
  https://en.wikipedia.org/wiki/Biological_database
- **List of biological databases (Wikipedia)** — Biological databases are stores of biological information. The journal Nucleic Acids Research regularly publishes special issues on biological databases and has a list of such databases. The 2018 issue has a list of about 180 such databases and updates to previously described databases.
  https://en.wikipedia.org/wiki/List_of_biological_databases
- **UniProt — about** — UniProt's own overview of the resource: the consortium behind it, its data sources, components and release cycle.
  https://www.uniprot.org/help/about
- **What is ENA?** — EMBL-EBI online course page introducing the European Nucleotide Archive: what it stores (raw reads, assemblies, annotated sequences), how it is organised and how it relates to GenBank and DDBJ.
  https://www.ebi.ac.uk/training/online/courses/ena-quick-tour/what-is-ena/
- **The UniProt databases** — EMBL-EBI online course page explaining the parts of UniProt: UniProtKB (Swiss-Prot and TrEMBL), UniParc, UniRef and the proteome sets.
  https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/the-uniprot-databases/
- **About RefSeq** — NCBI's description of the Reference Sequence (RefSeq) collection: curated, non-redundant reference sequences for genomes, transcripts and proteins, and how they differ from the GenBank archive.
  https://www.ncbi.nlm.nih.gov/refseq/about/
- **The NAR online Molecular Biology Database Collection** — The online catalogue of the annual Nucleic Acids Research database issue, listing well over a thousand molecular biology databases by category.
  https://www.oxfordjournals.org/nar/database/c/

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

2.13 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

2.14 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapter "Biological Data, Databases, and Provenance" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

2.14.1 Archival, derived and integrated resources

Distinguish the archival, curated, derived and integrated roles of biological databases. Explain why these roles are not a ranking of how reliable the data are.

Extensions, solution outline and common errors

Extensions:

- Classify the different kinds of information in UniProtKB, RefSeq and Ensembl, rather than giving each resource a single label.
- What happens when an integrated database copies an annotation from another integrated database?

Solution outline:

1. Define the labels as roles: archival (primary) resources preserve submitted observations; curated and derived resources select, transform, predict or interpret them; integrated resources link many kinds of entities and evidence.
2. One resource can play several roles. UniProtKB stores sequences translated from the nucleotide archives, computes automatic annotation (TrEMBL), curates the literature (Swiss-Prot) and integrates cross-references.
3. Contrast submission to GenBank, ENA and DDBJ with the selection and curation of RefSeq and the computed, integrated genome annotation of Ensembl.
4. "Primary" is not a quality label: archived data can contain submitter errors and contamination, while derived data add value but also assumptions.
5. Trust depends on tracing the specific claim back through its evidence, transformations and versions. Copying between resources does not create independent support.

Common errors:

- Treating "archival" as meaning "experimentally proven".
- Forcing every resource into one category.
- Counting three database pages that show the same annotation as three independent observations.

Connections: Section 2.2; annotation by similarity (Section 1.4); evidence for protein function (Section 10.4).

2.14.2 Identifiers, versions and mappings

Explain why a gene symbol is not enough for reproducible bioinformatics, and describe how identity can be preserved across resources and releases.

Extensions, solution outline and common errors

Extensions:

- Distinguish an accession from a versioned accession.
- Map an Ensembl transcript to a UniProt entry while keeping one-to-many results.
- Explain why a genomic coordinate is not a stable identifier.

Solution outline:

1. Gene symbols are labels for people. They can be ambiguous, have aliases, be reused or change, and an identifier means something only within its namespace.
2. A stable accession identifies a record; the version identifies one particular sequence. Record the database release and the retrieval date as well.
3. A coordinate needs the assembly, the contig accession, the coordinate convention (0- or 1-based) and the allele representation. Liftover between assemblies is a transformation, not an identity.
4. Treat mapping as a typed join: source accession and version, the relation, target accession and version, organism, and the source and release of the mapping.
5. Keep one-to-many and many-to-one mappings, find out whether they reflect isoforms, paralogues, merges or splits, and choose a rule that fits the question. Never silently take the first result.
6. Report how many identifiers mapped multiply or not at all.

Common errors:

- Removing version suffixes without recording it.
- Assuming that the "canonical" transcript is the dominant one in every tissue.
- Using gene symbols as keys in spreadsheets.
- Treating every cross-reference as an equivalence.

Connections: Section 2.2; gene models and assembly versions (Section 2.4, Section 5.6); summing transcripts to genes in RNA-seq (Section 7.3).

2.14.3 Selecting sequence and protein resources

You need a protein sequence for a functional analysis. How would you choose it, and justify the choice in terms of the biological entity, the evidence and the version?

Extensions, solution outline and common errors

Extensions:

- Why is the first BLAST hit not enough?
- Compare GenBank, RefSeq, UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, UniParc, UniRef and InterPro.
- State exactly what "reviewed" does and does not guarantee.

Solution outline:

1. Start from the biological entity: organism or strain, gene, isoform, and whether the submitted, the reference or an experimentally characterised sequence is needed.

2. Use the INSDC archives for submitted nucleotide records and their provenance, and RefSeq or Ensembl for versioned reference gene, transcript and protein models.
3. Use UniProtKB for protein sequences and functional annotation. Swiss-Prot entries are manually reviewed; TrEMBL entries are annotated automatically.
4. Use UniParc to follow the history of a sequence, UniRef when redundancy-reduced clusters are useful, and InterPro for family and domain signatures, which are predictions, not proof of biochemical activity.
5. "Reviewed" means that a curator has examined the entry, not that every statement has been demonstrated experimentally for that exact protein or isoform; check the evidence for each statement.
6. Check organism, sequence version, isoform, evidence, domain coverage and date, and record the accession and retrieval date.

Common errors:

- Treating sequence similarity as enough evidence to transfer a function.
- Assuming that a protein family can be defined by one identity threshold.
- Mixing a gene-level claim with an isoform-specific sequence.

Connections: Section 2.3; orthology and paralogy (Section 3.4); search statistics (Section 1.4); function prediction (Section 10.4).

2.14.4 Structure databases and evidence

A database shows a three-dimensional structure for a protein. What biological conclusions does the record support?

Extensions, solution outline and common errors

Extensions:

- Distinguish the PDB archive from portals such as RCSB PDB and PDBe.
- Compare experimental and predicted structures.
- Does a high resolution establish biological relevance?

Solution outline:

1. Establish the provenance: the PDB, run by the wwPDB, is the archive of experimentally determined structures, and RCSB PDB, PDBe and PDBj give access to it with their own added services. AlphaFold DB and similar resources contain computed models.
2. Compare the sequence of the construct with the biological protein: mutations, truncations, tags, missing residues, isoform, ligands and oligomeric state.
3. Check the quality measures for the method: resolution, R-free and geometry for crystallography; local resolution and model–map fit for cryo-EM; agreement with restraints for NMR.
4. Distinguish the asymmetric unit from the biological assembly.
5. For a predicted model, check the model version and its confidence (pLDDT per residue, PAE for the relative position of domains). A prediction is not experimental evidence of a conformation, an interaction or ligand binding.
6. The quality of the coordinates and their biological relevance are separate questions.

Common errors:

- Calling an AlphaFold model an experimental structure.
- Choosing the highest-resolution entry without checking the construct and state.
- Assuming that every ligand in a structure is physiological.
- Interpreting missing coordinates as missing sequence.

Connections: Section 2.5; structure determination and prediction (Section 9.4, Section 9.8); complexes (Chapter 11).

2.14.5 Variant databases

A sequencing study finds a human variant. Describe how databases can be used to assess it, keeping apart the evidence that it has been observed before, its population frequency, its molecular consequence and its clinical interpretation.

Extensions, solution outline and common errors

Extensions:

- Compare the questions answered by dbSNP, gnomAD and ClinVar.
- How should absence from gnomAD and conflicting ClinVar submissions be interpreted?
- Why are assembly and transcript versions needed?

Solution outline:

1. Keep the original call, with the reference assembly, contig, position, alleles and representation convention.
2. Use dbSNP (or study archives) to find identifiers and earlier observations, not pathogenicity.
3. Use gnomAD for population frequency, ancestry, coverage and gene constraint. Absence is weak evidence if coverage is poor or the relevant population is under-represented.
4. Use ClinVar for submitted interpretations of the variant for a specific condition, and check review status, submitter, criteria, date, mode of inheritance and conflicts.
5. Map the variant to versioned transcripts and protein isoforms; different transcripts can give different consequences.
6. Databases support, but do not replace, a clinical classification that combines segregation, phenotype, functional data and population evidence (for example under the ACMG/AMP guidelines).

Common errors:

- Taking an `rs` number as evidence that a variant is benign.
- Calling gnomAD participants healthy controls.
- Counting votes among ClinVar classifications.
- Mixing GRCh37 and GRCh38 coordinates.

Connections: Section 2.9; variant calling (Section 5.4); predicting the effects of mutations (Section 10.6).

2.14.6 Ontologies and enrichment

Explain why an ontology is a better computational representation than an unstructured list of biological terms, and what follows for functional enrichment analysis.

Extensions, solution outline and common errors

Extensions:

- Explain how GO annotations propagate in the graph, and the role of evidence codes.
- May all edges of an ontology be treated the same way?
- How do the ontology release and the choice of background affect an enrichment analysis?

Solution outline:

1. A controlled vocabulary standardises labels; an ontology adds stable identifiers, definitions and typed relations that can be used in computation.
2. A GO annotation links a gene product to a term, with an evidence code, a reference and sometimes qualifiers.
3. By the true path rule, an annotation to a term implies its ancestors, but only along relations that allow this inference; negations (NOT) must not be propagated.
4. Evidence codes show how an annotation is supported (experimental, phylogenetic, computational or electronic) and allow analyses to be repeated with stricter evidence.
5. Terms and the graph change between releases; record the ontology and annotation releases.
6. In enrichment analysis, the background should be the genes that could have been measured or selected, not every annotated gene, and annotation bias towards well-studied genes must be kept in mind.

Common errors:

- Treating GO terms as independent labels.
- Propagating annotations over every edge.
- Excluding electronic annotations without discussing what is lost.
- Using the whole genome as background for an assay that measured only part of it.

Connections: Section 2.6; gene-set analysis (Section 7.5); multiple testing (Section 7.4); evaluating function prediction (Section 10.7).

2.14.7 FAIR and controlled human data

Do the FAIR principles require all biomedical data to be openly downloadable? Describe how controlled-access human genomic data can be FAIR.

Extensions, solution outline and common errors

Extensions:

- What is the role of a data access committee?
- What obligations does a researcher take on when given access to data in dbGaP or EGA?

Solution outline:

1. FAIR means findable, accessible, interoperable and reusable: rich, machine-readable metadata, persistent identifiers, standard protocols and vocabularies, provenance and a clear licence.
2. "Accessible" can mean accessible after authentication and authorisation. FAIR is not the same as open, and it is not a privacy or ethics framework.

3. Controlled-access archives make the metadata public, while a data access committee decides whether a proposed use is compatible with the participants' consent and the restrictions on the data set.
4. Approved users must follow the data use agreement: secure storage, use only by approved people for the approved purpose, no attempts at re-identification, reporting of incidents, and deletion at the end of the project.
5. Data protection law (in Europe the GDPR) and ethical approval apply throughout the analysis and publication.

Common errors:

- Equating FAIR with free of charge or open.
- Assuming that de-identified data cannot be re-identified.
- Moving controlled data to an unapproved computer or cloud service.

Connections: Section 2.2, Section 2.9; human expression cohorts (Section 2.8); responsible use of data in machine learning (Section 12.7).

2.14.8 Reproducible programmatic retrieval

Describe a reproducible way to retrieve all records that match a biological criterion from an online database, keeping the provenance and detecting incomplete retrieval.

Extensions, solution outline and common errors

Extensions:

- What should be recorded about a query?
- How can an incomplete download be detected?
- Why should web pages not be scraped?

Solution outline:

1. Define the entity, the population, the type of evidence and the exclusion criteria before choosing the resource.
2. Use a fixed release or bulk download where possible; otherwise record the endpoint, date, literal query, filters, fields and API version.
3. Use the documented API, send requests in batches, follow the usage limits (for example the request rates of NCBI's E-utilities) and log errors.
4. Check that the number of records retrieved equals the number reported, handle pagination, and look for duplicates; keep the raw responses.
5. Check the biology: organism, assembly, sequence versions and controlled terms. Keep the mapping from raw to processed identifiers and all exclusions.
6. Archive the query record, the code and its environment, and cite the sources. For a live resource, repeat a sample of the query later to see how much it changes.

Common errors:

- Reporting only a download date.
- Scraping a web interface that may change at any time.
- Ignoring pagination and missing values.
- Assuming that a checksum proves that the data are biologically correct.
- Including controlled-access data in a public reproducibility package.

Connections: Section 2.2; data leakage and validation in machine learning (4).

2.14.9 Coordinates, formats and genome builds

A collaborator sends a BED file of enhancer peaks on hg19 and asks you to intersect it with variants from a VCF file on GRCh38 and with a GENCODE GTF annotation. List the checks you need, and the errors that can occur.

Extensions, solution outline and common errors

Extensions:

- A single-base variant at position 1,000 of chromosome 7 is written how in VCF and in BED?
- How is the length of an interval computed in each convention?
- Why may some peaks fail to lift over, or map to several places?

Solution outline:

1. Bring everything to the same genome build: lift the peaks from hg19 to GRCh38 with chain files (UCSC liftOver or CrossMap), and record how many fail or split; for important data, re-map the original reads instead.
2. Harmonise chromosome names (`chr7` vs 7) and check that contigs and patches match.
3. Handle coordinates: BED is 0-based, half-open (`chr7 999 1000`, length $e-s$), while VCF and GTF are 1-based, closed (position 1000, length $e-s+1$); use tools that know the conventions (BEDTools) rather than ad hoc arithmetic.
4. Check strand and feature type (gene, transcript, exon, CDS) in the GTF, and which annotation release was used.
5. Verify on a few known examples in a genome browser before trusting the full intersection.

Common errors:

- Mixing builds without noticing, because coordinates on both are plausible numbers.
- Off-by-one shifts that move a variant into the neighbouring base or out of a one-base feature.
- Silently getting zero overlaps because of chromosome naming.

Connections: Section 2.4, Section 2.2; whole-genome alignment and chains (Section 1.8); variant calling (Section 5.4).

3 Evolution

> **From the exam description**
>
> How does evolution affect the current genomes? Gene duplication, evolutionary pressure.

3.1 Overview

Dobzhansky's remark that "nothing in biology makes sense except in the light of evolution" is particularly true for bioinformatics. Almost every method that transfers information from one sequence to another assumes common descent, and sequence data are in turn our richest source of information about evolution. This chapter is largely historical: it follows how the main ideas of molecular evolution developed, and connects them with what has been learned from the analysis of genomes (Chapter 6) and protein structures (Chapter 9).

- the origin of life: prebiotic chemistry, the RNA world, metabolism-first scenarios and astrobiology (Section 3.2);
- the basic concepts: mutation, selection, drift, effective population size, hitchhiking, the neutral theory and the molecular clock (Section 3.3);
- homologues, orthology inference, orthology databases and gene duplication (Section 3.4); other routes to new genes, including *de novo* gene birth, and pseudogenes and small proteins are covered in Section 6.8;
- protein evolution: domains, repeats, disorder, coevolution, alternative splicing, structural constraints and ancestral sequence reconstruction (Section 3.5);
- detecting selection with dN/dS and related tests, and testing molecular convergence (Section 3.6);
- genome evolution: horizontal gene transfer, the power laws of gene family sizes, and gene-family expansion and contraction (Section 3.7).

Phylogeny, the reconstruction of evolutionary trees, including gene trees versus species trees, reconciliation and the tree of life, has its own chapter (Chapter 4).

3.2 The origin of life

All the processes described in this chapter assume a population of replicating organisms. How the first such system arose from chemistry is one of the great open questions in science. There is no agreed answer, but experiments, the geological record and comparative genomics constrain the possibilities. This section gives a short overview, starting with the history of the question. It helps to keep two kinds of statement apart. Some are established observations: amino acids and nucleobases form under plausible prebiotic conditions, RNA can catalyse reactions, and the catalytic centre of the ribosome is made of RNA. Others are hypotheses that try to connect these observations into a history, such as the RNA world, metabolism first or an origin in hydrothermal vents. None of the hypotheses has been demonstrated.

3.2.1 From spontaneous generation to prebiotic chemistry

For most of history it was assumed that life arose continually from non-living matter (*spontaneous generation*), for example maggots from meat. Experiments by Francesco Redi in the 17th century and Louis Pasteur in the 1860s showed that this does not happen under present conditions. The question then became how life arose once, under very different conditions on the early Earth.

In the 1920s Alexander Oparin and J. B. S. Haldane independently proposed that the early atmosphere contained no oxygen and was *reducing*, and that energy from lightning or ultraviolet light could drive the formation of organic molecules, which accumulated in the oceans as a "primordial soup". In 1953 Stanley Miller, a student of Harold Urey, tested the idea. The **Miller–Urey experiment** passed electric sparks through a mixture of methane, ammonia, hydrogen and water, and within a week produced several amino acids (Miller 1953). A reanalysis of Miller's stored samples with modern methods in 2008–2011 found more than 20 amino acids (Parker et al. 2011). Soon after, Joan Oró showed that adenine forms from ammonium cyanide (Oró 1960), so the building blocks of both proteins and nucleic acids could be made abiotically. Geochemists later concluded that the early atmosphere was probably less reducing than Miller assumed, which gives lower yields, but organic molecules could also have formed near volcanoes and hydrothermal vents, or been delivered by meteorites: the Murchison meteorite, which fell in Australia in 1969, contains dozens of amino acids.

3.2.2 When did life appear?

The Earth formed about 4.54 billion years ago. Its early history included a giant impact that formed the Moon, and possibly a period of heavy bombardment, but liquid water was present by about 4.4 billion years ago. The oldest generally accepted fossils, stromatolites (layered mats built by microbes) and microfossils, are about 3.5 billion years old. Older evidence, such as isotopically light carbon in a 4.1 billion-year-old zircon crystal (Bell et al. 2015) or possible microfossils in hydrothermal vent deposits that are at least 3.77 billion years old (Dodd et al. 2017), is debated. Molecular clocks calibrated with fossils place the last universal common ancestor (LUCA) at around or before 4 billion years ago, with wide uncertainty (Betts et al. 2018; Moody et al. 2024). Life therefore probably appeared early, soon after the Earth became habitable. LUCA was already a complex cell with a genetic code, ribosomes and hundreds of protein families; its reconstruction is discussed in Section 4.5. Everything between the first chemistry and LUCA must be inferred indirectly.

3.2.3 The RNA world

Modern life depends on a chicken-and-egg cycle: DNA stores the information, proteins catalyse its replication, and the instructions for making proteins are in DNA. In the 1960s Carl Woese, Francis Crick and Leslie Orgel independently suggested a way out: RNA can both store information and, in principle, catalyse reactions, so the first life could have used RNA for both. The idea became much more plausible when Thomas Cech and Sidney Altman discovered catalytic RNAs, **ribozymes**: a self-splicing intron (Kruger et al. 1982) and the RNA subunit of ribonuclease P (Guerrier-Takada et al. 1983). Walter Gilbert named the hypothesis the **RNA world** in 1986 (Gilbert 1986).

Several observations are consistent with it, although none proves it (Joyce 2002; Orgel 2004; Higgs and Lehman 2014):

- the ribosome is a ribozyme: the peptidyl transferase centre, where peptide bonds are formed, consists only of RNA (Ban et al. 2000), so protein synthesis itself is catalysed by RNA;
- many coenzymes, such as ATP, NAD, FAD and coenzyme A, contain nucleotides, perhaps as relics of RNA catalysis;

- deoxyribonucleotides are made from ribonucleotides in all organisms, consistent with DNA being a later invention;
- *in vitro* evolution has produced ribozymes that catalyse many reactions, including RNA ligases that replicate each other exponentially (Lincoln and Joyce 2009).

The main problem is how RNA itself could have formed. Ribose is unstable, and joining ribose to the bases is inefficient. In 2009 Powner, Gerland and Sutherland showed that activated pyrimidine ribonucleotides can form by a different route from simple precursors, without free ribose (Powner et al. 2009). This led to *cyanosulfidic* scenarios in which hydrogen cyanide, hydrogen sulfide and ultraviolet light produce precursors of nucleotides, amino acids and lipids in the same environment (Sutherland 2015). Most researchers today see the RNA world as a stage, not necessarily the first one, and many think RNA and peptides coevolved from the start. Whether an RNA world existed at all remains a hypothesis.

3.2.4 Metabolism first and the role of compartments

An alternative view is that self-sustaining networks of chemical reactions came before genetic information. Günter Wächtershäuser proposed that metabolism began on the surfaces of iron–sulfur minerals, driven by the formation of pyrite (Wächtershäuser 1988). William Martin and Michael Russell developed a related scenario in alkaline **hydrothermal vents**, where natural proton and redox gradients across thin mineral walls could have driven carbon fixation, and where the mineral pores could have served as the first compartments (Martin and Russell 2003). This is consistent with one reconstruction of LUCA as an anaerobic organism that used hydrogen and the acetyl-CoA pathway (Weiss et al. 2016) (a reconstruction that is itself debated), and with the fact that all life uses proton gradients to make ATP. These are scenarios, not demonstrated pathways: parts of such pathways have been reproduced with metal and mineral catalysts, but a complete, self-sustaining metabolic cycle without enzymes has not been shown, and critics point out that metabolic networks without a genetic system have limited ability to evolve. A third component that all scenarios need is a **compartment**: fatty acids form vesicles spontaneously, and in laboratory experiments such protocells can grow and divide, keeping replicating molecules together so that selection can act on them.

3.2.5 Astrobiology

Astrobiology studies the origin, evolution and distribution of life in the universe (Des Marais et al. 2008). It combines the origin-of-life research described above with the search for habitable environments and signs of life elsewhere. Since Michel Mayor and Didier Queloz discovered the first planet orbiting a Sun-like star in 1995 (Mayor and Queloz 1995), thousands of exoplanets have been found, many of them in the *habitable zone* where liquid water could exist. Within the Solar System, the main targets are Mars, which had liquid water early in its history, and the ice-covered oceans of Jupiter's moon Europa and Saturn's moon Enceladus. Astrobiology also studies *extremophiles* on Earth, organisms that live at high temperature, pressure, acidity or salinity, as models for life elsewhere. Two ideas from the field are relevant for biology on Earth. *Panspermia* proposes that life, or its building blocks, was transported between planets by meteorites. It does not solve the origin problem, but moves it elsewhere. The search for *biosignatures*, chemical or isotopic traces that only life produces, has also sharpened the debate about how to recognise the earliest life on Earth.

See also: further explanation in Section D.3.1; related Wikipedia articles in Section E.3.1.

Sources

Parts of this section are adapted from the Wikipedia articles Abiogenesis, Miller–Urey experiment, RNA world, Iron–sulfur world hypothesis, Earliest known life forms and Astrobiology (CC BY-SA 4.0, retrieved 3 October 2026).

3.3 Basic evolutionary concepts

Evolution at the molecular level is the result of a few processes acting together: mutation creates variation, and selection and random genetic drift decide which variants are lost and which spread through the population until they are *fixed*. Much of bioinformatics can be read as the analysis of the traces these processes leave in sequences. This section summarises the concepts and the history behind them.

3.3.1 A short history

Darwin's *On the Origin of Species* (1859) proposed descent with modification by natural selection, but without a theory of heredity. Mendel's laws (1866, rediscovered in 1900) provided the particulate inheritance that Darwin lacked. In the 1920s and 1930s Ronald Fisher, J. B. S. Haldane and Sewall Wright combined the two into **population genetics**, a mathematical theory of how allele frequencies change under mutation, selection, drift and migration. This was the core of the *modern synthesis* of the 1940s.

Molecular data entered in the 1950s and 1960s, when protein sequences such as haemoglobin, cytochrome *c* and fibrinopeptides could be compared between species. Two observations surprised many evolutionists: there was far more variation within species than expected, and proteins seemed to change at a roughly constant rate. These observations led to the neutral theory and the molecular clock (below), and to the **neutralist–selectionist debate** of the 1970s and 1980s. Today, genome-scale data are analysed with the same population-genetic framework, which provides the null models for detecting selection (Section 3.6).

3.3.2 Mutation

A **mutation** is any heritable change in the DNA sequence. The main kinds are:

- **point substitutions**, divided into *transitions* (purine ↔ purine, pyrimidine ↔ pyrimidine), which are about twice as frequent as *transversions*, partly because of the deamination of methylated cytosines in CpG dinucleotides;
- **insertions and deletions** (indels), often caused by replication slippage in repeats;
- **larger rearrangements**: duplications, deletions, inversions and translocations of chromosome segments, and the movement of transposable elements;
- duplication of whole chromosomes or genomes (Section 3.4).

In a protein-coding gene, a substitution can be **synonymous** (silent), changing the codon but not the amino acid, or **non-synonymous** (missense or nonsense). This distinction is the basis of the dN/dS test for selection.

Mutation rates are low. In humans there are about $1–1.5 \times 10^{-8}$ new point mutations per site per generation, or roughly 60–70 new mutations in each child, most of them from the father. Per generation and per genome, rates differ much less between species than per site, and Lynch has argued that the mutation rate itself evolves until further improvement in replication fidelity is too small to be seen by

selection, the *drift-barrier hypothesis* (Lynch et al. 2016). Mutations are not uniformly distributed: rates vary with sequence context, replication timing, chromatin state and position in the genome.

3.3.3 Selection

Natural selection changes allele frequencies because carriers of different alleles leave different numbers of descendants. Its strength is measured by the **selection coefficient** s, the relative difference in fitness. Three kinds are important in molecular evolution:

- **Purifying (negative) selection** removes harmful mutations. It is by far the most common, and it is why functionally important sequences are conserved, the basis of homology search (Chapter 1).
- **Positive (directional) selection** increases the frequency of beneficial mutations, which may then sweep to fixation.
- **Balancing selection** keeps several alleles in the population, for example through heterozygote advantage (sickle-cell haemoglobin in malaria regions) or frequency-dependent selection (immune genes such as the MHC).

3.3.4 Genetic drift and effective population size

In a finite population, allele frequencies also change by chance, because each generation is a random sample of the gametes of the previous one. This is **genetic drift**. In the simplest model (Wright–Fisher, with N diploid individuals), a new neutral mutation has a probability $1/(2N)$ of eventually becoming fixed, and when it does, fixation takes on average about $4N$ generations. Drift is therefore strong in small populations and weak in large ones.

The N in these formulas is not the census number of individuals but the **effective population size** N_e: the size of an idealised population that would show the same amount of drift. N_e is reduced by fluctuations in population size (bottlenecks), unequal sex ratios and variation in offspring number, and it is usually far smaller than the census size. Long-term estimates are of the order of 10^4 for humans, 10^5–10^6 for *Drosophila*, and 10^8 or more for many bacteria (Charlesworth 2009).

The interplay between selection and drift is governed by the product $N_e s$. When $|N_e s| \ll 1$, a mutation behaves as if it were neutral: drift decides its fate. When $|N_e s| \gg 1$, selection dominates. The same mutation can thus be effectively neutral in a small population and clearly selected in a large one. This is the key idea behind both the nearly neutral theory and Lynch's explanation for why genome complexity differs between prokaryotes and multicellular eukaryotes (Lynch and Conery 2003; Lynch et al. 2016) (Chapter 6).

The fixation probability of a selected mutation also depends on $N_e s$ (Kimura's diffusion approximation). In the simplest case, a new beneficial mutation with a small advantage s in a large population is fixed with a probability of only about $2s$: a mutation with a 1% advantage is lost by drift in about 98% of cases. The fixation probability of a slightly harmful mutation falls steeply once $N_e|s|$ exceeds about one. Because N_e differs by orders of magnitude between lineages, so does the efficacy of selection. In lineages with small N_e, such as large mammals or endosymbiotic bacteria, more slightly harmful mutations are fixed, which is seen, for example, as higher genome-wide dN/dS (Section 3.6) and weaker codon usage bias.

3.3.5 Genetic hitchhiking and linked selection

Selection on one site also affects nearby neutral sites, because they are inherited together until recombination separates them. When a beneficial mutation sweeps to fixation, linked neutral variants are dragged along: **genetic hitchhiking**, a term coined by Maynard Smith and Haigh in 1974 (Smith and Haigh 1974). A *selective sweep* reduces genetic diversity in the surrounding region, and the size of the region depends on the local recombination rate. The removal of harmful mutations has a similar but weaker effect on linked sites, called **background selection**. Because of these effects, levels of diversity along a genome are not set by drift alone, and regions of low recombination tend to have low diversity. Patterns of reduced diversity and skewed allele frequencies are used to find recent sweeps in population genomic data.

3.3.6 The neutral theory and the molecular clock

In 1968 Motoo Kimura (Kimura 1968), and independently Jack King and Thomas Jukes in 1969 (King and Jukes 1969), proposed the **neutral theory of molecular evolution**. It holds that most of the differences between species at the molecular level, and most of the variation within species, are caused not by positive selection but by the random drift of mutations that are selectively neutral. Harmful mutations occur but are quickly removed; beneficial ones are rare. The theory does not deny that adaptation happens. It claims that adaptive changes are a small minority of all molecular changes.

The neutral theory makes a simple prediction. The rate at which neutral mutations are *substituted* (fixed) in a lineage is the number of new mutations per generation, $2N\mu$, times their probability of fixation, $1/(2N)$:

$$k = 2N\mu \cdot \frac{1}{2N} = \mu.$$

The substitution rate equals the neutral mutation rate and does not depend on population size. If the mutation rate and the fraction of neutral mutations stay constant, neutral sequences therefore diverge at a constant rate. These conditions hold only approximately (see below), but the result explains why functionally less constrained sequences (synonymous sites, pseudogenes, introns, fibrinopeptides) evolve fastest: a larger fraction of their mutations are neutral.

The observation that came first was the **molecular clock**. In 1962 Zuckerkandl and Pauling noted that the number of amino acid differences between haemoglobins of different species increases roughly linearly with the time since divergence estimated from fossils (Pauling and Zuckerkandl 1963). Margoliash made the same observation for cytochrome *c*: all mammalian sequences are about equally different from all bird sequences, and both are equally different from fish. The clock made it possible to date divergences that left no fossils, and to build trees from sequences alone (Chapter 4). Tests showed that the clock is only approximate (Langley and Fitch 1974). Rates differ between genes, between sites and between lineages, for example with generation time and metabolic rate, and they can change over time within a lineage; rates measured over short time spans are often higher than those measured over long ones, partly because slightly harmful mutations have not yet been removed (Ho et al. 2005). A **strict molecular clock**, a single constant rate, is therefore a model to be tested, not a general law. Modern dating methods use **relaxed clocks** that allow the rate to vary across the tree, calibrated with several fossils.

Tomoko Ohta extended the theory in 1973 with the **nearly neutral theory** (Ohta 1973). Many mutations are only slightly harmful, with $|s|$ close to $1/N_e$. In small populations these behave as neutral and can be fixed by drift; in large populations they are removed. The theory explains why protein clocks tick at roughly constant rates per *year* rather than per generation: species with long generation times tend to have small populations, in which more slightly harmful mutations are fixed, and the two

effects partly cancel. Today the neutral and nearly neutral theories serve as the **null hypothesis** of molecular evolution. Adaptive explanations are invoked when the data deviate from neutral expectations (Section 3.6).

See also: further explanation in Section D.3.2; related Wikipedia articles in Section E.3.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Mutation, Natural selection, Genetic drift, Effective population size, Genetic hitchhiking, Neutral theory of molecular evolution, Nearly neutral theory of molecular evolution and Molecular clock (CC BY-SA 4.0, retrieved 3 October 2026).

3.4 Homologues, orthologues and gene duplication

3.4.1 Kinds of homology

The definitions were introduced in 1 and are only summarised here. Two genes are **homologous** if they descend from a common ancestral gene; homology is a yes-or-no statement, not a degree of similarity. Homologues are classified by the event that separated them (Fitch 1970, 2000; Koonin 2005):

- **orthologues** diverged by *speciation*;
- **paralogues** diverged by *gene duplication*. Relative to a given speciation, **in-paralogues** arose by duplication *after* it (and are together co-orthologous to genes in the other species), while **out-paralogues** arose *before* it (Remm et al. 2001);
- **xenologues** have a history that includes *horizontal gene transfer*;
- **ohnologues** are paralogues created by whole-genome duplication.

These definitions describe only the event that separated two genes. They say nothing about function: an orthologue can change function and a paralogue can keep it.

Similar features that evolved independently are **analogous** (convergent); the classic molecular example is the catalytic triad of chymotrypsin and subtilisin, which share a mechanism but not an ancestor. Convergence also occurs at the level of sequence, but on a small scale. Short linear motifs of a few residues (Section 3.5) arise independently many times, and the same amino acid replacements can be selected independently in different lineages, for example substitutions in the Na^+/K^+-ATPase that make insects feeding on toxic plants resistant to cardiac glycosides (Zhen et al. 2012). A shared short motif, or a few identical residues at functional sites, is therefore weak evidence of homology. Homology is inferred from significant similarity over a longer region (1), and even then convergent changes at a few sites can mislead tree inference.

The distinction between orthologues and paralogues matters for function prediction. A widely used heuristic, the *ortholog conjecture*, holds that orthologues are more likely to keep the same function than paralogues, which are freer to diverge after duplication. It underlies the transfer of functional annotation between species (Chapter 10), but it is a statistical tendency, not part of the definition. It has been tested and debated. One study using Gene Ontology annotations concluded that paralogues in the same species are *more* similar in function than orthologues (Nehrt et al. 2011); a reanalysis showed that this was largely a bias in how annotations are made, and that orthologues are on average somewhat more similar (Altenhoff et al. 2012). The difference is smaller than often assumed.

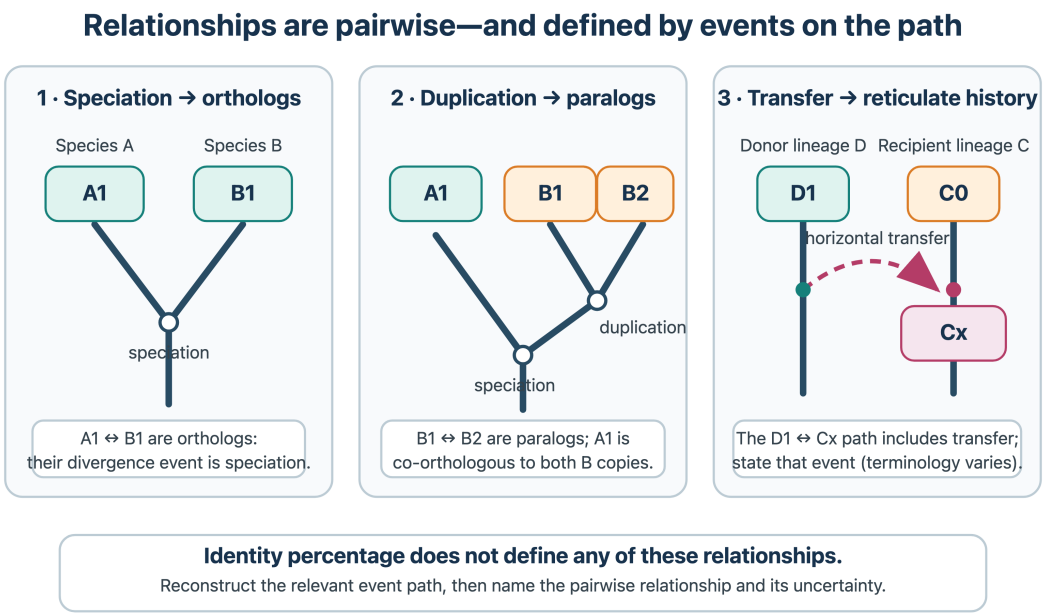


Figure 3.1: Homology relationships are defined by the event at which two genes diverged, not by their similarity. Genes separated by a speciation are orthologues, genes separated by a duplication are paralogues, and a gene in one species can be co-orthologous to two in-paralogues in another. A history that includes horizontal transfer gives xenologues; because that terminology is used less consistently, the transfer itself should be described. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

3.4.2 Inferring orthology

Orthology is defined by the gene tree and the species tree, but building and reconciling trees for every gene family is expensive. Methods therefore fall into two groups (Altenhoff et al. 2019):

Graph-based methods compare all proteins of all genomes with each other and cluster them using sequence similarity.

- **Reciprocal best hits (RBH).** Genes *a* in genome A and *b* in genome B are predicted to be orthologues if *b* is the best hit of *a* in B and *a* is the best hit of *b* in A. RBH is simple and fast and gives *candidate* one-to-one orthologues; it does not establish orthology by itself. It cannot handle in-paralogues (one gene in one genome is orthologous to several in the other). It fails when the true orthologue has been lost in one genome, because the best hit is then a paralogue, and when rates differ after a duplication, so that a slowly evolving paralogue scores higher than a fast-evolving orthologue.
- **COGs** (Clusters of Orthologous Groups), introduced by Tatusov, Koonin and Lipman in 1997, extend RBH to many genomes. Groups are built from triangles of reciprocal best hits in three genomes, which are merged when they share sides (Tatusov et al. 1997). COGs and their successors (KOGs, arCOGs, eggNOG; Huerta-Cepas et al. (2018)) are widely used for functional annotation of prokaryotic genomes.
- **InParanoid**, from Erik Sonnhammer's group in Stockholm, starts from reciprocal best hits between two genomes as "seed" orthologues, and then adds in-paralogues: genes in the same genome that are more similar to the seed than to any gene in the other genome (Remm et al. 2001).
- **OrthoMCL** clusters a graph of reciprocal best hits and in-paralogues from many genomes with the Markov clustering algorithm (Li et al. 2003).

Tree-based methods build a gene tree for each family and reconcile it with the species tree (Section 4.4). Duplication nodes separate paralogues and speciation nodes orthologues. Examples are Ensembl Compara, PhylomeDB and PANTHER. They are more accurate in principle, especially for complex families with many duplications and losses, but slower and sensitive to errors in the gene trees.

OrthoFinder (Emms and Kelly 2019) combines the two: it first clusters genes into *orthogroups* (all genes descended from one gene in the last common ancestor of the species considered) with a graph method, then infers a gene tree for each orthogroup, a rooted species tree from all of them, and finally the orthologues, paralogues and duplication events from the reconciled gene trees.

Methods disagree considerably, so the **Quest for Orthologs** consortium has developed shared reference proteomes and standardised benchmarks (Altenhoff et al. 2016). These show a trade-off between precision and recall: methods that find more true orthologues also report more false ones.

3.4.3 Orthology is not transitive

Orthology is a relationship between *pairs* of genes, and it is not transitive. If a is orthologous to b and b to c, a and c need not be orthologues. The simplest case is a pair of in-paralogues: if a gene duplicated in the human lineage after the split from mouse, both human copies are orthologous to the single mouse gene, but they are paralogues of each other. The same happens across species. Suppose a gene duplicated in the common ancestor of species A and C, after that lineage had separated from species B, and A and C then lost different copies. The gene a in A and the gene c in C are both orthologous to the gene b in B, but they diverged at the duplication, so they are paralogues, and a tree built from them would mislead (hidden paralogy; Section 4.4).

Relationships are therefore reported as *one-to-one*, *one-to-many* or *many-to-many*, and groups of orthologues in many species have to be defined relative to a point in the species tree. A **hierarchical orthologous group** (HOG) is the set of genes descended from a single gene in the last common ancestor of a chosen clade (Altenhoff et al. 2013). The orthogroups of OrthoFinder are HOGs at the root of the species analysed. A HOG defined at a deep level, for example vertebrates, contains all the copies made by later duplications, and splits into several HOGs at a shallower level, such as mammals, if the gene duplicated in between. A group of orthologues is therefore meaningful only together with its taxonomic level. The COGs and the deepest eggNOG groups are broad groups of this kind, and can contain ancient paralogues with different functions (Gabaldón and Koonin 2013).

3.4.4 Orthology databases

Most users do not infer orthologues themselves but take them from a database. The main resources differ in method, species and output:

- **OMA** (Roth et al. 2008; Altenhoff et al. 2020) infers pairwise orthologues from evolutionary distances estimated from all-against-all Smith–Waterman alignments. It removes hidden paralogues with *witnesses of non-orthology*: if a third genome contains two genes, one closer to each of the candidates a and b, the candidates probably descend from an older duplication, and their true orthologues have been lost. OMA reports pairwise orthologues, *OMA groups* (sets of genes, at most one per species, that are all orthologous to each other, which are conservative and useful as markers for phylogenomics) and HOGs.
- **OrthoDB** (Kuznetsov et al. 2022) provides HOGs at every major node of the taxonomy for many thousands of eukaryotes, prokaryotes and viruses. Its single-copy groups are the basis of the BUSCO sets used to assess the completeness of genome assemblies (Section 5.3).
- **eggNOG** (Huerta-Cepas et al. 2018) provides groups at many taxonomic levels together with functional annotations, which eggNOG-mapper transfers to new sequences (Chapter 10).

- **Tree-based resources** build a gene tree for each family and reconcile it with the species tree. *Ensembl Compara* (Vilella et al. 2008) does so for vertebrates and other eukaryotes in Ensembl, and assesses each predicted orthologue also by conserved gene order and whole-genome alignment. *PANTHER* (Mi et al. 2018) combines family trees with curated functional annotations, *TreeFam* (Li 2006) covered animal families, and *PhylomeDB* (Huerta-Cepas et al. 2013) stores the complete collection of gene trees (the *phylome*) of each genome.
- **Extensions of InParanoid** from Stockholm include Hieranoid, which applies InParanoid hierarchically along the species tree (Schreiber and Sonnhammer 2013), and Domainoid, which infers orthology for individual domains, so that proteins with rearranged domain architectures can be handled (Persson et al. 2019).
- **Meta-methods** such as MetaPhOrs combine the predictions of several phylogeny-based databases and score their consistency (Pryszcz et al. 2010). Agreement between databases is reassuring, but they are not independent if they share proteomes, trees or algorithms.

Databases disagree for several reasons that do not imply carelessness: different species sets and genome releases, the choice of one protein isoform per gene, gene models that are split or fused, thresholds, tree methods and the taxonomic level of the groups. A practical consequence is to record which resource, release and relationship type were used, and to check important cases with a gene tree, the domain architecture and conserved gene order (Section 6.5).

3.4.5 Gene duplication

Gene duplication is the main source of new genes, a view championed by Susumu Ohno in a 1968 paper (Ohno et al. 1968) and his 1970 book *Evolution by Gene Duplication.* Duplications arise in several ways:

- **unequal crossing-over** between misaligned chromosomes produces *tandem* duplications, which explains clusters of related genes such as the globins, olfactory receptors and Hox genes;
- **retrotransposition** of an mRNA produces an intronless copy (*retrocopy*) elsewhere in the genome, which usually lacks a promoter and becomes a processed pseudogene, but occasionally becomes a functional *retrogene* (Section 6.8);
- **segmental duplications** copy larger chromosomal regions;
- **whole-genome duplication** (polyploidy) copies everything at once.

Duplication is not the only source of new genes: genes also arise by fusion and fission, by horizontal transfer, by domestication of transposable elements and *de novo* from non-coding sequence. These mechanisms, together with the types of pseudogenes and the small proteins that classical annotation missed, are described in Section 6.8.

Duplications are frequent. Lynch and Conery estimated from genome sequences that a typical gene duplicates at a rate of the order of 0.01 per gene per million years, comparable to the point-mutation rate per site, but that most copies are lost again within a few million years (Lynch and Conery 2000).

3.4.5.1 The fates of duplicated genes

Immediately after duplication, the two copies are at least partly redundant, so the selective constraints on one or both of them may be relaxed, although both can also remain under purifying selection (Innan and Kondrashov 2010).

- **Nonfunctionalisation.** The usual fate is that one copy accumulates disabling mutations and becomes a pseudogene, which eventually decays beyond recognition.

- **Neofunctionalisation.** Ohno's classic model: one copy keeps the ancestral function, and the other acquires a new, beneficial function by mutation. The evolution of new enzyme specificities, or of the colour vision opsins in primates, are examples.
- **Subfunctionalisation.** In the *duplication–degeneration–complementation* model of Force and colleagues (Force et al. 1999), an ancestral gene with several functions or expression domains is duplicated, and each copy loses a different subset by degenerative mutations. Both copies are then needed to perform the original set of functions, and both are preserved, without any new function. Subfunctionalisation can be followed by specialisation of each copy.
- **Dosage effects.** Some duplicates are kept because more gene product is beneficial (for example rRNA genes). Others are sensitive to the balance between interacting proteins: duplicating one subunit of a complex can be harmful, which is why such genes are rarely retained after small duplications but often kept together after whole-genome duplication, when the balance is preserved (Papp et al. 2003).

3.4.5.2 Gene conversion and concerted evolution

Duplicated genes do not always evolve independently. **Gene conversion** is the non-reciprocal copying of a stretch of sequence from one gene onto a similar one during recombinational repair (Chen et al. 2007). Between paralogues it *homogenises* the copies, so that members of a family stay more similar to each other within a species than to their orthologues in other species. This is called **concerted evolution**; the tandem arrays of ribosomal RNA genes are the classic example. For sequence analysis, gene conversion has two consequences. Converted paralogues look younger than they are, which distorts the dating of duplications and can make paralogues appear as lineage-specific in-paralogues in orthology inference. Different parts of a gene can also have different histories, which violates the assumption of a single tree (Section 4.4). GC-biased gene conversion, which favours G and C alleles in heterozygotes, also affects base composition and can mimic positive selection.

3.4.5.3 Whole-genome duplication

Ohno also proposed that the vertebrate genome had been duplicated twice early in vertebrate evolution, providing the raw material for vertebrate complexity: the **2R hypothesis** (Ohno et al. 1968). It was controversial for decades because the signal had been eroded by gene loss and rearrangement. Genome sequences eventually settled the question. In yeast, Wolfe and Shields found duplicated blocks of genes in *Saccharomyces cerevisiae* and proposed an ancient genome duplication (Wolfe and Shields 1997). Kellis and colleagues proved it by comparison with *Kluyveromyces waltii*, which had not duplicated: each region of *K. waltii* corresponded to two regions in *S. cerevisiae*, and in most cases only one of the two copies of each gene had been kept (Kellis et al. 2004). For vertebrates, the analysis of the distribution of paralogous regions across the human genome provided strong support for two rounds of duplication in the ancestor of jawed vertebrates (Dehal and Boore 2005). Further duplications occurred in the ancestor of teleost fish (3R) and again in salmonids, and polyploidy is extremely common in plants, where most lineages show several ancient duplications.

After a whole-genome duplication most duplicates are lost, but the retained ohnologues are strongly biased towards genes involved in signalling, transcriptional regulation and development, consistent with the dosage-balance hypothesis. The duplicated *Hox* clusters of vertebrates are a classic example.

See also: further explanation in Section D.3.3; related Wikipedia articles in Section E.3.3.

Sources

Parts of this section are adapted from the Wikipedia articles Gene duplication, Subfunctionalization, Neofunctionalization and Paleopolyploidy (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Orthology: Concepts, Algorithms, and Databases" (CC BY 4.0), used with permission.

3.5 Protein evolution

Genes evolve, but selection acts on what the protein does. The structure, stability and interactions of a protein therefore shape how its sequence can change. This section collects what has been learned about protein evolution from comparisons of sequences, structures and genomes.

3.5.1 Domain evolution: combination, fusion and fission

Most proteins, especially in eukaryotes, consist of several domains, and domains are the units of protein evolution (Section 1.9). New proteins are mostly made by rearranging existing domains rather than by inventing new ones. Comparisons across genomes show several regularities (Apic et al. 2001; Björklund et al. 2005):

- The number of distinct domain *combinations* is small compared with what would be possible, and most domains have only one or two partners. A few promiscuous domains (for example SH3, SH2, PH and immunoglobulin domains) combine with very many others (Basu et al. 2008).
- Domain architectures change mainly by gaining or losing domains at the N- or C-terminus, rather than by insertions in the middle (Björklund et al. 2005).
- **Gene fusion** joins domains that previously formed separate proteins, and **gene fission** splits them. Both occur (Snel et al. 2000), but in a comparison of more than a hundred genomes fusion was about four times as frequent as fission (Kummerfeld and Teichmann 2005). Because fused domains usually work together, fusion events are used to predict functional links between proteins that are separate in other species (the "Rosetta stone" method; Chapter 10).
- Multidomain proteins are more common in eukaryotes than in prokaryotes, and animals have invented many new domain combinations, particularly in signalling and cell adhesion.
- The order in which subunits of a protein complex assemble often mirrors the order in which they were added during evolution (Levy et al. 2008).

3.5.2 Repeat proteins

Many proteins contain **tandem repeats**: the same unit, from a few residues to a whole domain, repeated many times. Repeats of short motifs of about 20–40 residues often stack into elongated *solenoid* structures, such as ankyrin repeats, leucine-rich repeats (LRR), tetratricopeptide repeats (TPR) and HEAT/armadillo repeats. Others close into rings, such as the β-propellers of WD40 proteins. Larger repeats consist of whole domains, for example the immunoglobulin and fibronectin domains of titin. Kajava classified repeat proteins by the length of the unit and the resulting structure (Kajava 2012).

Repeats evolve quickly by internal duplication and loss of units, through unequal crossing-over or replication slippage. Comparative analyses show that repeated domains are much more common in eukaryotes, especially in animals, and that repeat number often differs even between close homologues (Björklund et al. 2006). Because each unit adds a binding surface, repeat proteins are well suited for protein–protein interactions, and their modular nature has made them popular scaffolds for protein

design. For sequence analysis, repeats are a challenge: they cause ambiguous alignments, inflated scores and difficulties in defining homologous positions.

3.5.3 Evolution of intrinsically disordered proteins

Many proteins, or long regions of them, do not fold into a single stable structure. These **intrinsically disordered regions** (IDRs) are common in eukaryotes, where a substantial fraction of all residues is predicted to be disordered, and rarer in prokaryotes (Lee et al. 2014). They often act as flexible linkers, or bind other proteins through **short linear motifs** (SLiMs) of a few residues, as in many signalling and regulatory proteins.

Disordered regions evolve differently from folded domains (Brown et al. 2011):

- Their sequences change much faster, and insertions and deletions are frequent. When homologous proteins differ in length, the extra residues are mostly found in disordered regions, which suggests that disordered regions are the main place where proteins can grow or shrink (Light et al. 2013).
- What is conserved is often not the sequence but *properties*: the amino acid composition, net charge, the fact of being disordered, or the presence of short motifs at varying positions. Such regions are therefore hard to align, and standard substitution matrices and alignment methods, designed for folded domains, fit them poorly.
- Short motifs can appear and disappear quickly by a few mutations, which makes disordered regions a rapid source of new interactions and regulation.

3.5.4 Coevolution

Residues that interact in a structure cannot change independently. If a mutation at one position destabilises a contact, a compensating mutation at the interacting position can restore it. Over many homologues, this produces **correlated mutations** between positions in a multiple sequence alignment. The idea was tested in the 1990s (Göbel et al. 1994), but simple correlation measures gave poor predictions of contacts because they cannot separate direct couplings from indirect, transitive correlations. Global statistical models, notably **direct coupling analysis** (DCA), largely solved this problem (Weigt et al. 2009; Marks et al. 2011). Given deep enough alignments, the strongest couplings predict residue contacts accurately, which made the first accurate *de novo* structure predictions of large proteins possible and is part of the information used by AlphaFold (Chapter 9). The same principle applied to concatenated alignments of two interacting protein families predicts which proteins interact and how (Cong et al. 2019) (Chapter 11). Coevolution is also the basis of comparative RNA structure analysis, where it appears as compensatory base pair changes (Section 1.7).

3.5.5 Alternative splicing and its evolution

In eukaryotes, the exons of one gene can be combined in different ways, producing several mRNA and protein **isoforms**. RNA-seq showed that almost all human multi-exon genes, about 95%, are alternatively spliced (Pan et al. 2008; Wang et al. 2008). Comparisons between species show that splicing patterns evolve fast: within the same tissue, splicing differs more between species than gene expression does, and many alternative exons are specific to one lineage (Barbosa-Morais et al. 2012; Merkin et al. 2012). Alternatively spliced segments often lie in disordered regions and contain binding motifs, so tissue-specific splicing can rewire protein interactions (Buljan et al. 2012).

How much of this is functional is debated. Proteomics experiments detect only a small fraction of the predicted isoforms, most genes have one dominant isoform, and alternative isoforms are often poorly conserved. This suggests that much alternative splicing is noise from an imprecise splicing machinery,

which is tolerated because it has little effect (Tress et al. 2017). Alternative splicing is therefore an example of a general question in evolutionary genomics: whether a widespread molecular activity is an adaptation or a by-product (Chapter 6).

3.5.6 Structural and functional constraints on protein evolution

Comparing sequences with structures has shown that the rate of evolution varies systematically, both between sites in a protein and between proteins (Echave et al. 2016).

- **Between sites.** Buried residues in the hydrophobic core evolve slowly, and surface residues fast. The rate of a site correlates with its relative solvent accessibility and, even more, with how many contacts it makes (*packing density*). Active sites and binding interfaces are conserved, which is how functional sites are recognised in alignments. The exchanges allowed at a site depend on its environment: hydrophobic for hydrophobic in the core, and much freer exchanges on the surface.
- **Between proteins.** The strongest predictor of how fast a protein evolves is not its functional importance but its **expression level**: highly expressed proteins evolve slowly (Drummond et al. 2005). One explanation is that mistranslated or misfolded copies of an abundant protein are costly to the cell, so there is strong selection for sequences that fold robustly.
- **Stability.** Most proteins are only marginally stable, and most mutations are destabilising. Neutral evolution under a requirement for minimal stability alone reproduces this marginal stability and the robustness of proteins to mutation (Taverna and Goldstein 2002). A destabilising but beneficial mutation can often only be accepted after a stabilising one, which makes evolutionary paths depend on the order of mutations (*epistasis*).
- **Structure is conserved longer than sequence.** Homologous proteins keep the same fold long after sequence similarity has disappeared, and the structural divergence of the core increases slowly as identity falls (Chothia and Lesk 1986). This is why structure is used to detect remote homology (1), and why the number of distinct folds in nature is limited (Holm and Sander 1996).

3.5.7 Ancestral sequence reconstruction

The same substitution models used to build trees (Section 4.3.6) can be used to infer the sequences of extinct ancestors at the internal nodes of a tree. Early methods used parsimony; modern methods use maximum likelihood or Bayesian inference. Given an alignment, a tree and a model, they compute for each site and each node the posterior probability of every amino acid or nucleotide (Yang et al. 1995; Williams et al. 2006). The reconstructed gene can then be synthesised and the "resurrected" protein studied in the laboratory. This has been used to test hypotheses about how new functions and specificities arose, which mutations were needed and in which order (epistasis), the temperature of ancient environments from the thermostability of ancestral proteins, and as a starting point for protein engineering (Thornton 2004).

A reconstruction is an estimate, not an observation, and its uncertainty should be taken into account:

- The posterior probabilities are conditional on the alignment, the tree and the substitution model. Errors in any of them propagate, and insertions and deletions are usually handled crudely.
- Uncertainty grows with the depth of the node and with the rate of the site. Deep ancestors may have many sites where two or more residues are almost equally likely.
- The common practice of taking the most probable residue at every site is biased. It systematically picks the most frequent and often stabilising residues, so such reconstructions tend to overestimate stability. Robustness is therefore tested by also making alternative ancestors with the second most probable residues, or by sampling sequences from the posterior distribution, and checking that the conclusions do not change.

See also: further explanation in Section D.3.4; related Wikipedia articles in Section E.3.4.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Tandem repeat, Intrinsically disordered proteins, Short linear motif and Alternative splicing (CC BY-SA 4.0, retrieved 3 October 2026).

3.6 Detecting selection

The neutral theory provides a null model (Section 3.3): if a sequence evolves neutrally, all kinds of mutations are fixed at the rates at which they occur. Selection shows up as a deviation from this expectation. For protein-coding genes, the most widely used comparison is between two kinds of sites in the same gene.

Selection acts on organisms: individuals with different phenotypes leave different numbers of offspring. The methods in this section do not observe this. They detect *statistical signatures* in sequence data that deviate from a neutral null model, and they attribute the deviation to selection under the assumptions of that model. A statement such as "this residue is under positive selection" is shorthand for "a codon model assigns this site to a class with $\omega > 1$ with high posterior probability". It is an inference that can be wrong if the model is wrong, and it says nothing directly about which phenotype was selected.

3.6.1 dN/dS (Ka/Ks)

Synonymous mutations do not change the protein and are, to a first approximation, neutral. Non-synonymous mutations change an amino acid and are exposed to selection on the protein. The ratio

$$\omega = \frac{d_N}{d_S} \quad \text{(also written } K_a/K_s\text{)}$$

compares the number of non-synonymous substitutions per non-synonymous site (d_N) with the number of synonymous substitutions per synonymous site (d_S) between two sequences or along a branch of a tree. Normalising *per site* is essential, because only about a quarter of random point mutations in coding sequence are synonymous. The interpretation is:

- $\omega < 1$: **purifying selection** predominates. Most amino acid changes are harmful and removed. This is the norm; typical genes have ω around 0.1–0.3, and highly constrained genes much lower.
- $\omega \approx 1$: consistent with **neutral evolution**, as in pseudogenes.
- $\omega > 1$: evidence for **positive (diversifying) selection** under the model. Amino acid changes have been favoured, for example in genes involved in immunity, host–pathogen conflicts or reproduction.

These are statements about an *average* over all the sites and branches compared. A gene-wide $\omega < 1$ does not mean that every site is under purifying selection, and it can coexist with positive selection at a few sites or on a few branches (next section). A value near one can also be an average of constrained and positively selected sites. Conversely, $\omega > 1$ is evidence of positive selection only if the model's assumptions hold: neutral synonymous sites, a correct alignment, and a single tree without recombination (see Pitfalls below).

The first methods estimated d_N and d_S by counting sites and differences between two sequences, with corrections for multiple substitutions (Nei and Gojobori 1986). Modern methods use **codon models**,

Markov models of substitution between the 61 sense codons in which ω is an explicit parameter, fitted by maximum likelihood on a phylogeny (Goldman and Yang 1994; Yang and Bielawski 2000).

3.6.2 Site, branch and branch-site models

Averaged over a whole gene and a whole tree, ω is almost always below one, even when positive selection has acted, because a few adaptive changes are swamped by the many conserved sites. Codon models therefore let ω vary:

- **Site models** allow ω to differ between codons. A model that includes a class of sites with $\omega > 1$ (for example M2a or M8) is compared with a nested model that does not (M1a or M7) by a likelihood ratio test, and Bayesian methods then identify the sites that most probably belong to the class with $\omega > 1$ (Nielsen and Yang 1998). A classic example is the peptide-binding groove of the MHC molecules.
- **Branch models** let ω differ between branches of the tree, for example to test whether a gene evolved faster after a duplication or in a particular lineage.
- **Branch-site models** test for positive selection on a few sites along a specified (*foreground*) branch (Zhang 2005).

These models are implemented in programs such as PAML (codeml) and HyPhy (Yang 2007; Pond et al. 2004).

3.6.3 Tests based on polymorphism and divergence

dN/dS was developed for comparing species. Within a population, polymorphisms include slightly harmful mutations that have not yet been removed, so ω computed between closely related sequences behaves differently and is hard to interpret (Kryazhimskiy and Plotkin 2008). The **McDonald–Kreitman test** (McDonald and Kreitman 1991) uses both kinds of data. Under neutrality, the ratio of non-synonymous to synonymous changes should be the same for polymorphisms within a species and for fixed differences between species. An excess of non-synonymous fixed differences indicates adaptive evolution, and its size can be used to estimate the fraction of amino acid substitutions driven by positive selection. McDonald and Kreitman first applied it to the alcohol dehydrogenase gene of *Drosophila*. Population genomic data are also scanned for the signatures of selective sweeps (Section 3.3): local reductions in diversity, skewed allele frequencies and long haplotypes.

3.6.4 Molecular convergence

The same amino acid change in independent lineages with a similar way of life is tempting evidence of adaptation. Parker and colleagues compared the genomes of echolocating bats and dolphins and reported convergent substitutions in many genes, including some involved in hearing (Parker et al. 2013). Reanalyses showed that comparable levels of convergence occur between lineages that do not echolocate, so that most of the signal was not specific to echolocation (Thomas and Hahn 2015; Zou and Zhang 2015). Convergence is common by chance because each site accepts only a few amino acids, and because related lineages share substitution tendencies.

A test of adaptive convergence therefore needs:

- a species tree and a reconstruction of ancestral states (Section 3.5), to show that the same *derived* state arose independently on the lineages concerned. A shared ancestral state, or a state shared through introgression or incomplete lineage sorting (Section 4.4), is not convergence;

- a null expectation, estimated from control branches matched for length or from simulations with a model that keeps site-specific amino acid preferences;
- functional evidence that the change affects the phenotype. The substitutions in the Na^+/K^+-ATPase of insects that feed on toxic plants are a convincing case because their effect on toxin resistance was tested experimentally (Zhen et al. 2012).

3.6.5 Pitfalls

Tests for selection are powerful but easy to misuse.

- **Saturation.** When sequences are very divergent, synonymous sites have changed many times and d_S cannot be estimated reliably. When they are very similar, there are too few changes for a meaningful test.
- **Alignment and annotation errors** create apparent non-synonymous changes and are a common source of false positives.
- **Recombination** violates the assumption of a single tree.
- **Synonymous sites are not always neutral.** Codon usage bias, splicing signals and mRNA structure constrain them, and biased gene conversion can mimic selection.
- **Multiple testing.** Genome-wide scans test thousands of genes and many branches, and need correction for multiple testing (Section 7.4.1).

See also: further explanation in Section D.3.5; related Wikipedia articles in Section E.3.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Ka/Ks ratio and McDonald–Kreitman test (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Molecular Evolution and Comparative Reasoning" (CC BY 4.0), used with permission.

3.7 Genome evolution

The structure of genomes, how many genes they contain, how they are organised and how they differ between the domains of life, is covered in Chapter 6. This section discusses evolutionary questions that genome sequences made it possible to answer.

3.7.1 Horizontal gene transfer

Horizontal (lateral) gene transfer (HGT) is the movement of genetic material between organisms other than from parent to offspring. In prokaryotes it occurs through three main mechanisms:

- **transformation**, the uptake of free DNA from the environment;
- **transduction**, transfer by bacteriophages;
- **conjugation**, transfer of plasmids or integrative elements through direct cell-to-cell contact.

HGT was first recognised through the spread of antibiotic resistance between bacterial species in the 1950s. Complete genome sequences showed how common it is (Ochman et al. 2000; Koonin et al. 2001). Transferred genes can be recognised in several ways:

- **phylogenetic incongruence**: the gene tree conflicts strongly with the species tree (Section 4.4);

- **phyletic patterns**: a gene is present in a species but absent from all its close relatives, and present in distantly related organisms;
- **composition**: recently transferred genes often have a GC content or codon usage that differs from the rest of the genome. This signal fades as the gene adapts to its new host.

Each method has false positives (for example, gene loss can mimic a patchy distribution), and estimates of the amount of HGT vary. It is clear, however, that in bacteria and archaea gene content changes rapidly. Strains of the same species can differ in a large fraction of their genes, which led to the concept of the **pan-genome**: a *core* of genes shared by all strains and an *accessory* part present in only some of them (Tettelin et al. 2005). Transfer, rather than duplication, is the main way prokaryotic gene families expand (Treangen and Rocha 2011). Transferred genes are often involved in adaptation to new niches: antibiotic resistance, virulence, metabolism of new compounds.

HGT challenges the idea of a single tree of life. Doolittle argued in 1999 that, if genes are transferred frequently, the history of prokaryotes is better described as a network than a tree (W. F. Doolittle 1999). In practice, a core of genes that are rarely transferred, such as those for ribosomal proteins, still gives a consistent tree that is used to classify prokaryotes (Section 4.5), but many other genes do not follow it. In eukaryotes, HGT is much rarer, with an important exception: the transfer of genes from the endosymbionts that became mitochondria and plastids to the nuclear genome (*endosymbiotic gene transfer*).

3.7.2 Power laws and the statistics of gene families

When the first complete genomes were compared, a striking regularity appeared. The number of protein families (or folds, or domains) of a given size n falls off approximately as a **power law**,

$$P(n) \propto n^{-\gamma},$$

rather than exponentially. In every genome, a few families are very large (protein kinases, ABC transporters, zinc fingers), and most families have only one or a few members. This was shown for gene families by Huynen and van Nimwegen (Huynen and Nimwegen 1998), and for domain and fold occurrence by Qian, Luscombe and Gerstein (Qian et al. 2001). The exponent differs between genomes: larger genomes have flatter distributions, with relatively more large families.

These distributions are interesting because they can be explained by simple models of evolution. In **birth–death–innovation models**, families grow by duplication of one of their members (with a probability proportional to the family size, as in "rich get richer" models), shrink by gene loss, and new families are created by innovation, such as horizontal transfer or the origin of new genes. Such models reproduce the observed distributions without assuming selection for particular family sizes (Koonin et al. 2002). A related regularity is that the number of genes in certain functional categories scales with genome size as a power law: the number of transcription regulators grows faster than linearly with the number of genes, and the number of genes for basic functions such as translation grows slower (Nimwegen 2003). Similar power laws are found in other biological networks (Chapter 11). Together with other statistical regularities, they have been called the "laws" of genome evolution (Koonin 2011).

The power laws also have practical consequences for bioinformatics. A few very large families dominate databases and search results, and an appreciable fraction of the genes in any newly sequenced genome belongs to small families or is unique to the lineage ("orphan" genes), for which homology-based annotation has nothing to offer (Chapter 10). Some orphans are genuinely new genes born *de novo* from non-coding DNA, while others are old genes whose homologues have diverged beyond detection; how the two are told apart is discussed in Section 6.8.

3.7.3 Gene-family expansion and contraction

A family with 20 members in one species and five in its relatives may have expanded in that lineage, but the counts alone do not show it. Family sizes change all the time by duplication and loss, so a count must be compared with what is expected along the species tree. The usual approach models family size as a **birth–death process** on a dated species tree: each copy duplicates at rate λ and is lost at rate μ, the likelihood of the observed counts at the leaves is computed, and families whose changes are improbable under the genome-wide rates are flagged. The program CAFE introduced this approach (Hahn et al. 2005; De Bie et al. 2006), and CAFE5 allows the rates to vary between families (Mendes et al. 2020). Applied to mammals, it showed rapid turnover of gene families: humans and chimpanzees differ by at least 6% of their genes in copy number (Demuth et al. 2006).

The analysis is only as good as the counts. Collapsed repeats in an assembly hide copies, while uncollapsed haplotypes, fragmented gene models and pseudogenes add false ones, and differences in assembly and annotation quality between genomes easily dominate the signal (Section 5.3). An unusual change in family size is a statement about history, not about adaptation. Linking an expansion to a phenotype requires further evidence, such as expression, biochemical divergence of the copies and repeated expansions in independent lineages with the same phenotype.

See also: further explanation in Section D.3.6; related Wikipedia articles in Section E.3.6.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Horizontal gene transfer and Pan-genome (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Molecular Evolution and Comparative Reasoning" (CC BY 4.0), used with permission.

3.8 Key concepts

A checklist for revision.

- **Origin of life:** established observations versus hypotheses; spontaneous generation and Pasteur; Oparin–Haldane; the Miller–Urey experiment and its reanalysis; the age of the Earth and the earliest evidence of life; the RNA world and the evidence for it (ribozymes, the ribosome, coenzymes); prebiotic synthesis of nucleotides; metabolism-first and hydrothermal-vent scenarios; protocells; astrobiology, exoplanets and biosignatures
- **Population genetics:** types of mutation and mutation rates; purifying, positive and balancing selection; genetic drift and fixation probability (neutral $1/(2N)$, beneficial about $2s$); effective population size and why $N_e s$ matters for the efficacy of selection; hitchhiking, selective sweeps and background selection
- **Neutral theory:** Kimura and King–Jukes; why the neutral substitution rate equals the mutation rate; the molecular clock (Zuckerkandl and Pauling, Margoliash) and its limits (rate variation between lineages, genes, sites and over time; strict versus relaxed clocks); Ohta's nearly neutral theory; the neutralist–selectionist debate
- **Homology:** orthologues, paralogues (in-/out-), xenologues, ohnologues, analogues; relationship type versus functional outcome; convergent sequence evolution of short motifs and single sites; the ortholog conjecture and the debate about it
- **Orthology inference:** reciprocal best hits as candidate one-to-one orthologues and their failure modes; COGs, InParanoid, OrthoMCL, eggNOG; tree-based methods; OrthoFinder; Quest for Orthologs benchmarks; orthology is not transitive; one-to-one, one-to-many and many-to-many

relationships; hierarchical orthologous groups and their taxonomic level; orthology databases (OMA and witnesses of non-orthology, OrthoDB, eggNOG, Ensembl Compara, PANTHER, PhylomeDB, Hieranoid, Domainoid, MetaPhOrs) and why they disagree
- **Gene duplication:** mechanisms (tandem, retrotransposition, segmental, whole-genome); non-, neo- and subfunctionalisation (DDC model); dosage balance; gene conversion and concerted evolution; whole-genome duplications in yeast, vertebrates (2R), fish and plants, and how they were proven
- **Other origins of new genes** (see Section 6.8): retrogenes, fusion and fission, HGT, TE domestication, *de novo* gene birth and proto-genes; orphan genes versus homology detection failure; pseudogenes as the usual fate of duplicates and retrocopies
- **Protein evolution:** domain combinations, fusion and fission; repeat proteins; how disordered regions evolve; coevolution and direct coupling analysis; alternative splicing and the noise-versus-function debate; site-specific rates and structure, expression level and rate, marginal stability and epistasis; ancestral sequence reconstruction and its uncertainty
- **Selection:** selection on fitness versus statistical signatures in sequences; dN/dS and its interpretation (an average over sites and branches; model assumptions); counting methods and codon models; site, branch and branch-site models (PAML, HyPhy); the McDonald–Kreitman test; molecular convergence, null models and the echolocation debate; pitfalls
- **Genome evolution:** mechanisms and detection of HGT; pan-genomes; tree versus network of life; power-law distributions of family sizes and birth–death–innovation models; scaling of functional categories with genome size; birth–death models of gene-family size (CAFE) and why an expansion is not evidence of adaptation

3.9 Recommended reading

These papers are recommended for reading in full:

- Kimura (1968) — the neutral theory in two pages; why most molecular change may be neutral.
- Koonin (2005) — orthologues, paralogues and evolutionary genomics; the definitions and why they matter.
- Lynch and Conery (2000) — how often genes duplicate and what happens to the copies.
- Echave et al. (2016) — why rates of evolution vary among protein sites; the link between structure and evolution.
- Treangen and Rocha (2011) — horizontal transfer, not duplication, drives gene family expansion in prokaryotes.
- Emms and Kelly (2019) — OrthoFinder; how orthology is inferred from gene trees at genome scale.

Titles, abstracts, links and the full texts are in Section A.3; a longer list of further papers is in Section B.3.

3.10 Background reading (open access)

- **Abiogenesis (Wikipedia)** — Abiogenesis or the origin of life is the natural process by which life arises from non-living matter, such as simple organic compounds.
  https://en.wikipedia.org/wiki/Abiogenesis
- **RNA world (Wikipedia)** — The RNA world is a hypothetical stage in the evolutionary history of life on Earth in which self-replicating RNA molecules proliferated before the evolution of DNA and proteins. The term also refers to the hypothesis that posits the existence of this stage.
  https://en.wikipedia.org/wiki/RNA_world

- **Astrobiology (Wikipedia)** — Astrobiology is a scientific field within the life and environmental sciences that studies the origins, early evolution, distribution, and future of life in the universe by investigating its deterministic conditions and contingent events.
  https://en.wikipedia.org/wiki/Astrobiology
- **Genome evolution (Wikipedia)** — Genome evolution is the process by which a genome changes in structure (sequence) or size over time. The study of genome evolution involves multiple fields such as structural analysis of the genome, the study of genomic parasites, gene and ancient genome duplications, polyploidy, and comparative genomics.
  https://en.wikipedia.org/wiki/Genome_evolution
- **Neutral theory of molecular evolution (Wikipedia)** — The neutral theory of evolution holds that most of the genetic variation within and between species is due to random genetic drift of mutant alleles that are selectively neutral. A neutral mutation is one that does not affect an organism's ability to survive and reproduce.
  https://en.wikipedia.org/wiki/Neutral_theory_of_molecular_evolution
- **Molecular clock (Wikipedia)** — The molecular clock is a figurative term for a technique that uses the mutation rate of biomolecules to deduce the time in prehistory when two or more life forms diverged. The biomolecular data used for such calculations are usually nucleotide sequences for DNA, RNA, or amino acid sequences for proteins.
  https://en.wikipedia.org/wiki/Molecular_clock
- **Genetic drift (Wikipedia)** — Genetic drift is the change in the frequency of an existing allele in a population due to chance.
  https://en.wikipedia.org/wiki/Genetic_drift
- **Genetic hitchhiking (Wikipedia)** — Genetic hitchhiking, also called genetic draft or the hitchhiking effect, is when an allele changes frequency not because it itself is under natural selection, but because it is near another gene that is undergoing a selective sweep and that is on the same DNA chain.
  https://en.wikipedia.org/wiki/Genetic_hitchhiking
- **Sequence homology (Wikipedia)** — Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life.
  https://en.wikipedia.org/wiki/Sequence_homology
- **Gene duplication (Wikipedia)** — Gene duplication is a mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.
  https://en.wikipedia.org/wiki/Gene_duplication
- **Intrinsically disordered proteins (Wikipedia)** — In molecular biology, an intrinsically disordered protein (IDP) is a protein that lacks a fixed or ordered three-dimensional structure, typically in the absence of its macromolecular interaction partners, such as other proteins or RNA.
  https://en.wikipedia.org/wiki/Intrinsically_disordered_proteins
- **Ka/Ks ratio (Wikipedia)** — In genetics, the Ka/Ks ratio, also known as ω or dN/dS ratio, is used to estimate the balance between neutral mutations, purifying selection and beneficial mutations acting on a set of homologous protein-coding genes.
  https://en.wikipedia.org/wiki/Ka/Ks_ratio
- **Horizontal gene transfer (Wikipedia)** — Horizontal gene transfer (HGT) or lateral gene transfer (LGT) is the movement of genetic material between organisms other than by the ("vertical") transmission of DNA from parent to offspring (reproduction). HGT is an important factor in the evolution of many organisms.
  https://en.wikipedia.org/wiki/Horizontal_gene_transfer
- **Pan-genome (Wikipedia)** — In the fields of molecular biology and genetics, a pan-genome is the entire set of genes from all strains within a clade. More generally, it is the union of all the genomes of a clade.

https://en.wikipedia.org/wiki/Pan-genome

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

3.11 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

3.12 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Evolution and Comparative Genomics" and "Orthology: Concepts, Algorithms, and Databases" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

3.12.1 Mutation, selection and drift

Explain how mutation, natural selection and genetic drift together shape molecular evolution.

Extensions, solution outline and common errors

Extensions:

1. What is the difference between a mutation and a substitution?
2. Why does the neutral substitution rate equal the mutation rate in an ideal model?
3. Does a larger population experience less drift?
4. What does "nearly neutral" mean?
5. Is mutation random?

Solution outline:

- Mutation creates heritable variants. It is not directed towards future usefulness, although its rate and spectrum vary between sites, sequence contexts and organisms.
- Selection changes the expected reproductive success associated with variants. Drift is random sampling in finite populations, and acts even when selection is present.
- A mutation is a new change; a substitution is a change that has become fixed along a lineage. For neutral mutations in a diploid population, $2N\mu$ new copies arise per generation and each has fixation probability $1/(2N)$, so the substitution rate is $K = \mu$ (Kimura 1968).
- Compare selection with drift through the product $N_e s$. A weakly deleterious mutation behaves as effectively neutral when its effect is small relative to drift, the central idea of the nearly neutral theory (Ohta 1973).
- The effective population size is usually much smaller than the census size, and differs between parts of the genome.

Common errors:

- Saying that drift acts only at neutral sites.

- Treating every fixed difference as adaptive.
- Saying that "random mutation" means that all mutations are equally likely.
- Treating $|N_e s| \approx 1$ as a sharp boundary rather than a rough scale.
- Confusing allele-frequency change within a population with substitution between lineages.

Connections: Section 3.3; branch lengths in Section 4.2; population variation in Section 6.6.

3.12.2 Sequence identity and orthology

Two proteins from different species are 40% identical. What can be inferred about homology and orthology, and what further evidence is needed?

Extensions, solution outline and common errors

Extensions:

1. Does similarity prove homology?
2. Can paralogues occur in different species?
3. What are co-orthologues?
4. How would you infer the relationship?
5. Does orthology imply the same function?

Solution outline:

- Identity is a property of an alignment. What it means depends on the length and coverage of the alignment, the composition of the sequences, their domains and a statistical significance estimate. Homology is common ancestry, and orthology is defined by the event that separated the genes.
- Orthologues diverged at a speciation and paralogues at a duplication. A duplication followed by a speciation gives paralogues in different species, and a duplication after a speciation makes two copies co-orthologous to one gene in the other lineage (Fitch 1970).
- Collect the gene family with suitable species and outgroups, align the homologous regions, infer a gene tree with support values, reconcile it with an independent species tree, and check conserved gene order and domain architecture.
- Report the relationship as one-to-one, one-to-many or many-to-many, with its uncertainty.
- Orthology supports, but does not prove, transfer of function; check conserved functional residues, expression and direct experimental evidence.

Common errors:

- Defining homology or a gene family by a universal identity threshold.
- Calling the proteins "40% homologous".
- Assuming that genes in different species are orthologues and genes in the same species are paralogues.
- Treating reciprocal best hits as proof of orthology after a duplication.
- Ignoring matches that cover only one domain.

Connections: 1; Section 3.4; gene trees and reconciliation in Section 4.4; annotation transfer in Section 10.4.

3.12.3 Inferring duplication and loss

A gene occurs twice in species A, once in species B, and is not found in species C. Suggest and compare plausible evolutionary histories.

Extensions, solution outline and common errors

Extensions:

1. Is one duplication plus one loss the answer?
2. Why is the species tree needed?
3. How does the choice of species change the reconstruction?
4. What does conserved synteny add?
5. How would you confirm the absence in species C?

Solution outline:

- Several histories fit the counts: a duplication in the lineage of A; an older duplication with different losses; horizontal transfer; or technical errors. The counts alone cannot decide between them.
- Check the assemblies and annotations, search the genome sequence itself, look for pseudogenes and uncollapsed haplotypes, and choose outgroups.
- Infer a gene tree and reconcile it with an independently supported species tree under a duplication–loss or duplication–transfer–loss model. Parsimony minimises a weighted number of events, while probabilistic methods model the rates of events.
- Use conserved gene order: the corresponding position in another genome can reveal a missed gene, or ancient duplicates after their sequences have diverged.
- Vary the species sampled, the tree topology, the family boundaries and the event costs or rates, and report ambiguous nodes rather than forcing one history.

Common errors:

- Treating the absence of an annotation as a biological loss.
- Inferring the history from a table of counts without a tree.
- Assuming that the history with the fewest events is the true one.
- Building the species tree from the gene of interest, and then treating the agreement as independent support.
- Ignoring whole-genome duplication and the subsequent loss of most duplicates.

Connections: Section 3.4; reconciliation in Section 4.4; synteny in Section 6.5; assembly and annotation in Section 5.3 and Section 5.6.

3.12.4 Interpreting d_N/d_S

A gene has an estimated $d_N/d_S = 1.4$. Does this support positive selection, and which analyses are needed before drawing that conclusion?

Extensions, solution outline and common errors

Extensions:

1. How are d_N and d_S estimated?

2. How do branch, site and branch-site models differ?
3. Can positive selection occur when the ratio for the whole gene is below one?
4. What if d_S is close to zero or saturated?
5. Why do recombination and GC-biased gene conversion matter?

Solution outline:

- The estimate indicates an excess of non-synonymous substitutions under its codon model, alignment and tree, averaged over sites and branches; on its own it is not sufficient.
- Codon models account for the genetic code, multiple substitutions, codon frequencies, the transition–transversion bias and the phylogeny (Goldman and Yang 1994).
- A ratio for the whole gene averages over sites and branches. Episodic selection at a few residues can coexist with purifying selection on the rest of the protein; branch-site models test this alternative.
- Test a pre-specified alternative against the appropriate null model with a likelihood-ratio test, check that the optimisation has converged, correct for multiple testing, and identify candidate sites only after the model-level test is significant.
- Check the codon alignment, non-homologous regions, sequencing errors, saturation of d_S, recombination, tree uncertainty, selection on synonymous sites, mutation bias and GC-biased gene conversion. Recombination can give false positives in site models (Anisimova et al. 2003).
- Confirm the functional effect experimentally, and use polymorphism data if the claim concerns recent adaptation.

Common errors:

- Calling $\omega = 1$ proof of neutrality.
- Interpreting $\omega < 1$ as absence of any positive selection.
- Dividing raw counts of differences without correcting for multiple substitutions.
- Treating the posterior probability of a single site as an independent hypothesis test.
- Equating selection on a coding sequence with a specific phenotype.

Connections: Section 3.6; substitution models and likelihood in Section 4.3.6; multiple testing in Section 7.4.1.

3.12.5 Testing gene-family expansion

A desert plant has 20 annotated copies of a gene family involved in water stress, while related plants have five. Design an analysis that distinguishes unusual copy-number evolution from an adaptive expansion.

Extensions, solution outline and common errors

Extensions:

1. Why is 20 against five not enough?
2. What does a birth–death model do?
3. How can assembly quality inflate the size of a family?
4. What evidence connects an expansion to adaptation?
5. How should the phylogeny enter an association with a trait?

Solution outline:

- Check that the copies are distinct loci, with read depth, flanking sequence, intact domains and complete gene models; separate pseudogenes, fragments, alleles and uncollapsed haplotypes.
- Define the family at an explicit evolutionary depth, sample species and outgroups, infer a gene tree and reconcile duplications and losses.
- Model copy number along a dated species tree. A birth–death model, as in CAFE, tests whether the change is unusual compared with the normal turnover of families; recent versions allow rates to differ between families (Mendes et al. 2020).
- Repeat the analysis with alternative family definitions and allowing for errors in the counts, and correct for testing many families.
- An expansion is a statement about history, not adaptation. Connect the copies to stress-responsive expression, biochemical differences, phenotypes after perturbation, and repeated expansions in independent lineages with the same ecology.

Common errors:

- Calling every BLAST hit a functional gene.
- Ignoring differences in annotation quality between the focal genome and its relatives.
- Equating a statistically unusual expansion with positive selection.
- Treating species as independent observations without correcting for phylogeny.
- Assuming that all duplicates share one function.

Connections: Section 3.7; gene duplication in Section 3.4; assembly errors in Section 5.3; expression in Chapter 7.

3.12.6 Detecting horizontal gene transfer

A metabolic gene occurs in one bacterium and in several distantly related archaea, but not in the close relatives of the bacterium. Assess the hypothesis of horizontal transfer.

Extensions, solution outline and common errors

Extensions:

1. Which signals would you combine?
2. Why can GC content fail?
3. Does conflict between gene tree and species tree prove transfer?
4. How do contamination and gene loss compete as explanations?
5. Can you infer the direction of transfer and the donor?

Solution outline:

- A patchy distribution motivates the hypothesis but can also result from gene loss, incomplete assemblies or missed annotations.
- Collect a broad, quality-controlled family, infer a gene tree with support values and reconcile it with the species tree under a duplication–transfer–loss model. Compare alternative roots and models.
- Inspect GC content, codon usage, mobile elements, the boundaries of the insertion, association with plasmids and the conservation of the neighbouring genes. Composition detects mainly recent transfers from compositionally different donors; the signal fades as the gene adapts to its new genome.
- Check the raw reads, coverage, the contig context and independent assemblies to exclude contamination.

- Infer the direction only when rooting, taxonomic distribution and genomic context support it. The true donor may be unsampled or extinct.
- Describe the transfer explicitly, since the terminology for xenologues is less uniform than for orthologues and paralogues.

Common errors:

- Using unusual GC content as proof.
- Calling every conflict between gene tree and species tree a transfer.
- Assuming that the closest sampled sequence comes from the actual donor.
- Ignoring incomplete lineage sorting and introgression in eukaryotes.
- Failing to rule out contamination.

Connections: Section 3.7; conflicting trees in Section 4.4; pan-genomes in Section 6.6; metagenomic binning in Section 5.7.

3.12.7 Testing molecular convergence

Two distantly related marine mammals share the same amino acid at one site of a protein. Formulate a test of the hypothesis that this is adaptive convergence.

Extensions, solution outline and common errors

Extensions:

1. What counts as an independent origin?
2. Which null model is appropriate?
3. Why reconstruct ancestral sequences?
4. Could introgression or ancestral polymorphism explain the pattern?
5. What functional evidence is needed?

Solution outline:

- First establish, with a species tree and probabilistic ancestral reconstruction, that the same derived state arose independently on the branches concerned. A shared ancestral state, or one shared through gene flow, is not convergence.
- Count or score the convergent substitutions and compare them with a null expectation that keeps the branch lengths, the substitution tendencies and, preferably, site-specific constraints. Matched control branches are more informative than arbitrary species without the phenotype.
- Consider alignment, tree, recombination, incomplete lineage sorting, introgression, mutation bias and uncertainty in the ancestral states.
- Show an excess over the expectation. In the echolocation debate, apparent genome-wide convergence depended strongly on the choice of control branches (Thomas and Hahn 2015).
- Link the substitution to a biochemical or organismal phenotype by mutagenesis or structural analysis, and ideally show it in several independent origins.
- State a graded conclusion: a repeated state; an excess of molecular convergence; an association with the phenotype; or an experimentally supported adaptive mechanism.

Common errors:

- Treating similar phenotypes as proof of the same molecular route.

- Counting species instead of independent evolutionary transitions.
- Using a null model that ignores site-specific amino acid preferences.
- Calling identical states adaptive without functional evidence.
- Ignoring gene flow.

Connections: Section 3.6; ancestral sequence reconstruction and structural constraints in Section 3.5; incomplete lineage sorting in Section 4.4.

4 Phylogeny

> **From the exam description**
>
> Advantages and limitations of different methods.

4.1 Overview

Phylogenetics reconstructs the evolutionary history of genes and species from their sequences. It is one of the oldest areas of bioinformatics, and many of its ideas, such as dynamic programming on trees, explicit probabilistic models of sequence change, the bootstrap and Markov chain Monte Carlo, have spread to the rest of the field. The problem is both computational, because the number of possible trees is astronomically large, and statistical, because the data are noisy and the models approximate. This chapter covers

- the history of the field and the vocabulary of trees: rooting, clades and the number of possible trees (Section 4.2);
- the main families of methods: distance methods, parsimony, maximum likelihood and Bayesian inference (Section 4.3), the substitution models they depend on, the choice between nucleotide, codon and protein data, and tests of the molecular clock (Section 4.3.6), and measures of support (Section 4.3.9);
- alignment uncertainty, phylogenomics, gene trees versus species trees, consensus trees and concordance factors, reconciliation, phylogenomic function prediction and phylogenetic networks (Section 4.4);
- what has been learned about the tree of life, its root and the origin of eukaryotes, how taxonomy relates to phylogeny, and about the evolution of humans and the other great apes (Section 4.5).

The evolutionary processes that trees describe are covered in Chapter 3, and the use of trees to infer orthology in Section 3.4.

4.2 Basic definitions

A phylogenetic tree is a hypothesis about how a set of sequences or species descend from common ancestors. Trees are used to classify organisms, infer orthology (Section 3.4), date divergences, reconstruct ancestral sequences, detect selection (Section 3.6) and track the spread of pathogens. This section introduces the history of the field and the vocabulary used for trees. The methods for building them follow in Section 4.3.

4.2.1 From morphology to molecules

Darwin's only figure in the *Origin of Species* (1859) was a branching tree, and Ernst Haeckel drew elaborate "trees of life" in the 1860s, but these were based on anatomy and intuition. Two developments in the 1950s and 1960s turned phylogeny into a quantitative science.

The first was **cladistics**. Willi Hennig argued that classification should reflect phylogeny only, and that groups should be *clades*, an ancestor and all its descendants (Hennig 1965). Clades are recognised by shared *derived* characters (synapomorphies), not by shared ancestral characters. Having a backbone, for example, does not show that fish form a group, because the backbone was inherited from the common ancestor of all vertebrates, including us. Hennig's ideas led to the first algorithmic methods, which searched for the tree that needed the fewest character changes (parsimony).

The second was the arrival of **molecular data**. Zuckerkandl and Pauling described molecules as "documents of evolutionary history", and noticed that proteins seemed to change at a roughly constant rate, the molecular clock (Section 3.3). In 1967 Fitch and Margoliash built one of the first trees from protein sequences, using cytochrome *c* from 20 species and a distance method (Fitch and Margoliash 1967). In the same years Cavalli-Sforza and Edwards introduced maximum likelihood and Jukes and Cantor the first probabilistic model of sequence change (Jukes and Cantor 1969). Joseph Felsenstein showed in 1978 that parsimony can be misleading (Felsenstein 1978) and in 1981 made maximum likelihood practical for DNA sequences (Felsenstein 1981). Neighbour joining followed in 1987 (Saitou and Nei 1987) and Bayesian methods in the late 1990s (Rannala and Yang 1996; Huelsenbeck et al. 2001). The history of these methods, and the debates between their proponents, are told in Felsenstein's book (Felsenstein 2004).

4.2.2 Tree terminology

A tree consists of **nodes** connected by **branches** (edges).

- The **leaves** (tips, terminal nodes) are the observed sequences or species, often called *taxa* or operational taxonomic units (OTUs).
- The **internal nodes** represent hypothetical ancestors. In a fully resolved (*bifurcating*) tree each internal node has two descendants. A node with more descendants is a **polytomy**. It can represent a genuine rapid radiation (a *hard* polytomy) or, more often, lack of resolution (a *soft* polytomy).
- The **topology** is the branching pattern alone. **Branch lengths** usually give the expected number of substitutions per site along the branch, that is, the amount of change, which is the product of time and rate. A branch length is therefore *not* a time, unless the tree has been dated with a clock model (a *chronogram*). In a time tree the branch lengths are proportional to time and the tree is *ultrametric*: all leaves are equally far from the root. In a *cladogram* only the topology is shown, and the branch lengths mean nothing.
- A **clade** (monophyletic group) is an ancestor and all its descendants. A **paraphyletic** group contains an ancestor and only some of its descendants (reptiles without birds, prokaryotes without eukaryotes). A **polyphyletic** group is defined by a character that evolved independently, without including the common ancestor (warm-blooded animals). Cladistic classification accepts only monophyletic groups. Clades are defined only on a rooted tree. An unrooted tree defines *splits* (bipartitions of the leaves), which become clades once a root is chosen.
- **Sister groups** are two clades that are each other's closest relatives: they descend from the same immediate ancestral node. The **outgroup** is a taxon known to lie outside the group of interest (the *ingroup*).

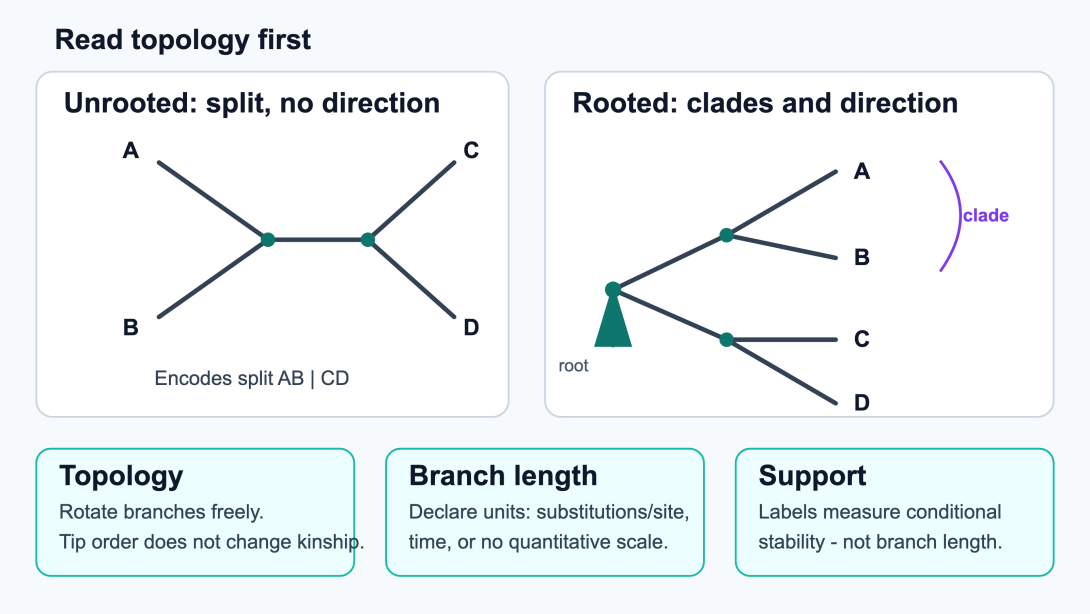


Figure 4.1: Reading a tree. An unrooted tree defines splits, such as AB|CD, but no direction in time; rooting it defines clades. Rotating the branches around a node does not change the topology, so the order of the tips on the page says nothing about relatedness. Branch lengths and support values are separate quantities, and the units of the branch lengths must be stated. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

4.2.2.1 Rooted and unrooted trees

A **rooted** tree has a node, the root, that represents the most recent common ancestor of all leaves, and so gives every branch a direction in time. An **unrooted** tree shows only the relationships between the leaves. Most reconstruction methods produce unrooted trees, because the substitution models they use are *time-reversible*: the likelihood is the same wherever the root is placed. A tree is rooted afterwards, usually by

- **outgroup rooting**: including a more distant taxon and placing the root on the branch that leads to it. This is the most reliable method, but it requires prior knowledge that the outgroup really lies outside the ingroup, and a distant outgroup can itself be attracted to long branches (Section 4.3);
- **midpoint rooting**: placing the root halfway along the longest path between two leaves. It needs no outgroup but assumes a roughly constant rate, and it misplaces the root when some lineages evolve faster than others;
- a **molecular clock** model, or non-reversible models, which put the root where the clock or the model fits best. These use information in the ingroup alone, but the rooting signal is often weak, so the root position is uncertain.

The root of the whole tree of life cannot be found with an outgroup, because no outgroup exists. How it has been approached is described in Section 4.5.

4.2.2.2 The number of possible trees

The main computational problem in phylogenetics is that the number of trees grows super-exponentially with the number of taxa. An unrooted bifurcating tree with n leaves has $2n - 3$ branches. A new leaf can be attached to any of them, so the number of unrooted trees is

$$U(n) = \prod_{i=3}^{n}(2i-5) = (2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!},$$

and the number of rooted trees is $R(n) = U(n+1) = (2n-3)!!$. There are 3 unrooted trees for four taxa, 15 for five, about 2×10^6 for 10 and more than 2×10^{20} for 20. For 50 taxa the number exceeds 10^{74}. Finding the best tree under the parsimony or likelihood criterion is NP-hard, so every practical method combines a criterion for scoring trees with a heuristic for searching among them (Yang and Rannala 2012; Kapli et al. 2020).

4.2.2.3 Writing trees down

Trees are stored as text in the **Newick format**, in which each clade is written within parentheses and branch lengths follow a colon. For example, `((human:0.006,chimp:0.007):0.002,gorilla:0.009,orangutan:0.` describes an unrooted tree of four apes. Newick was agreed upon informally in 1986 at a meeting in Newick's seafood restaurant in Durham, New Hampshire. Richer formats such as NEXUS (used by PAUP* and MrBayes) and the XML-based PhyloXML and NeXML can store data and annotations together with the tree.

4.2.3 Gene trees and species trees

A tree built from one gene is a **gene tree**. It describes the history of that gene, which need not be the same as the history of the species that carry it, the **species tree** (Maddison 1997). Incomplete lineage sorting, duplication and loss, horizontal gene transfer and recombination are distinct biological processes that make gene trees differ from species trees. In addition, an estimated gene tree can differ from the species tree simply because it was estimated with error. This distinction is central to phylogenomics, orthology inference and the study of human origins, and is discussed in Section 4.4.

See also: further explanation in Section D.4.1; related Wikipedia articles in Section E.4.1.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Phylogenetic tree, Cladistics, Monophyly, Phylogenetics and Newick format (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Phylogenetic Inference" (CC BY 4.0), used with permission.

4.3 Phylogenetic methods

Tree-building methods differ in two ways: whether they work on pairwise distances or directly on the alignment columns (characters), and whether they use an explicit statistical model of sequence evolution. The four main families are distance methods, maximum parsimony, maximum likelihood and Bayesian inference (Yang and Rannala 2012; Kapli et al. 2020). All of them start from a multiple sequence alignment (Section 1.6), and all of them assume that the alignment is correct and that sites evolve independently. The models the statistical methods depend on are described in Section 4.3.6, and measures of support in Section 4.3.9.

4.3.1 Distance methods

Distance methods first compute an $n \times n$ matrix of pairwise **evolutionary distances** and then build the tree that best fits the matrix. The observed proportion of differing sites, the p-distance, underestimates the true number of substitutions, because a site can change more than once (multiple hits, back mutations). It is therefore corrected with a substitution model. Under the Jukes–Cantor model (Jukes and Cantor 1969), for example,

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right),$$

which equals p for small p and goes to infinity as p approaches the saturation level of $3/4$. For proteins, distances are corrected with amino acid replacement matrices (Section 4.3.6). Because of saturation, distances between very divergent sequences have large errors.

UPGMA (unweighted pair group method with arithmetic mean) is the simplest clustering method, introduced by Sokal and Michener in 1958 for numerical taxonomy. It repeatedly joins the two closest clusters and defines the distance from the new cluster to every other as the average of the distances of its members. UPGMA always produces a rooted, ultrametric tree. It is correct only if the distances are themselves ultrametric, that is, if the sequences evolved approximately under a molecular clock: when lineages evolve at different rates, a fast-evolving taxon looks far from its true sister and is joined in the wrong place. For this reason it is rarely used for phylogenies today, although the same algorithm is used for guide trees in multiple alignment and for clustering in general.

Neighbour joining (NJ), introduced by Saitou and Nei (Saitou and Nei 1987), is based on a different principle and does not assume a clock. It requires only that the distances are *additive* (the distance between two leaves is the sum of the branch lengths on the path between them), so the two lineages that descend from a node may have changed by different amounts. At each step it does not join the closest pair, but the pair that minimises the total length of the tree. This is found with a corrected matrix

$$Q(i,j) = (r-2)\,d(i,j) - \sum_{k} d(i,k) - \sum_{k} d(j,k),$$

where r is the number of clusters left. Subtracting the total distance of each taxon from all others compensates for long branches: a taxon that is far from everything is not penalised for being on a long branch. The pair with the lowest Q is joined, the branch lengths from the new node to the two taxa are estimated, the distances from the new node to all other taxa are computed, and the procedure is repeated until three nodes remain (Studier and Keppler 1988). NJ produces an unrooted tree in $O(n^3)$ time, which must be rooted by one of the methods in Section 4.2. It is *consistent*: if the distances are additive (exactly fit a tree), it recovers that tree. It was later shown to be a greedy algorithm for the balanced minimum-evolution criterion (Gascuel 2006).

Minimum evolution chooses the tree with the smallest sum of branch lengths, with branch lengths fitted to the distances by least squares. FastME searches tree space under this criterion and is both fast and accurate (Desper and Gascuel 2002). The older **Fitch–Margoliash** method instead minimises the weighted squared differences between the observed distances and the distances along the tree (Fitch and Margoliash 1967).

Distance methods are fast and are still used for very large data sets, for starting trees in likelihood searches, and for guide trees. Their weakness is that reducing an alignment to pairwise distances discards information: which sites changed, and how changes are shared among several taxa.

4.3.2 Maximum parsimony

Maximum parsimony chooses the tree that explains the data with the fewest character changes, following Hennig's principle that shared derived characters indicate common descent. For a given tree, the minimum number of changes at a site is found with **Fitch's algorithm** (Fitch 1971), a dynamic programme over the tree. Going from the leaves to the root, each internal node is given a set of states: the intersection of the sets of its two children if it is non-empty, and otherwise their union, at the cost of one change. The total number of unions over all sites is the parsimony score of the tree. **Sankoff's algorithm** generalises this to weighted changes, for example to make transitions cheaper than transversions (Sankoff 1975). Only *parsimony-informative* sites, where at least two states each occur in at least two taxa, can distinguish between trees.

Scoring a tree is fast, but finding the best tree still requires a search over tree space. Exact *branch-and-bound* searches are possible for about 20 taxa; larger data sets need the heuristics described below. Parsimony was the dominant method from the 1970s to the 1990s, and PAUP* and TNT were the main programs.

Parsimony has no explicit model of evolution and gives no branch lengths in units of substitutions per site. Its main problem is that it is not **statistically consistent** in all conditions. Felsenstein showed that when two non-sister taxa have long branches and the others short ones, parallel changes on the long branches are more frequent than the true shared changes on the short internal branch (Felsenstein 1978). Parsimony then groups the long branches together, and more data only make it more confident in the wrong tree. This region of parameter space is called the *Felsenstein zone*, and the artefact **long-branch attraction** (LBA). LBA affects all methods when the model is wrong, but model-based methods can correct for it to a large extent by accounting for multiple hits. A classic example is the early placement of fast-evolving microsporidian parasites at the base of the eukaryotic tree; they were later shown to be related to fungi.

4.3.3 Maximum likelihood

Maximum likelihood (ML) evaluates each tree by the probability of the observed alignment given the tree T, its branch lengths and an explicit substitution model θ (Section 4.3.6):

$$L(T, \theta) = P(D \mid T, \theta) = \prod_{s=1}^{m} P(D_s \mid T, \theta),$$

where the product is over the m alignment columns D_s, assumed to evolve independently. The likelihood of one column is a sum over all possible states at the internal nodes, a number that grows exponentially with the number of taxa. **Felsenstein's pruning algorithm** (Felsenstein 1981) computes it efficiently by dynamic programming, in the same way as the forward algorithm of an HMM (Section 1.6). Let $L_k(x)$ be the probability of the data below node k, given that node k has state x. For a leaf, $L_k(x)$ is 1 for the observed state and 0 otherwise. For an internal node with children i and j at branch lengths t_i and t_j,

$$L_k(x) = \Big[\sum_{y} P_{xy}(t_i)\, L_i(y) \Big] \Big[\sum_{z} P_{xz}(t_j)\, L_j(z) \Big],$$

where $P_{xy}(t)$ is the probability that state x changes into y along a branch of length t. At the root, $P(D_s) = \sum_x \pi_x L_{\text{root}}(x)$, with π_x the equilibrium frequencies. The cost is linear in the number of taxa and sites and quadratic in the number of states (4 for DNA, 20 for proteins, 61 for codons). In practice, log-likelihoods are summed over sites to avoid numerical underflow.

For a fixed topology, the branch lengths and model parameters are optimised numerically. The topology is then improved by a **heuristic tree search** that starts from an initial tree (often from neighbour joining or parsimony) and applies rearrangements:

- **NNI** (nearest-neighbour interchange) swaps subtrees across an internal branch, giving $2(n-3)$ neighbours;
- **SPR** (subtree pruning and regrafting) removes a subtree and reattaches it to any other branch;
- **TBR** (tree bisection and reconnection) cuts the tree in two and reconnects the parts in all possible ways.

Larger rearrangements explore tree space more widely but are slower. Since the search can get stuck in local optima, programs use several starting trees, random perturbations and stochastic search. The most widely used programs are **RAxML** and its successor RAxML-NG (Stamatakis 2014), **IQ-TREE** (Minh, Schmidt, et al. 2020), which combines fast hill climbing with stochastic perturbations and integrates model selection and fast bootstrap, **PhyML** (Guindon et al. 2010), and **FastTree**, which gives approximate ML trees for alignments of up to millions of sequences (Price et al. 2010).

Maximum likelihood is statistically consistent when the model is correct, uses all the information in the alignment, and allows hypotheses to be tested with likelihood-ratio tests. Its disadvantages are its computational cost and its dependence on the model: a poorly fitting model can give a confidently wrong tree.

4.3.4 Bayesian inference

Bayesian inference uses the same likelihood but combines it with **prior** distributions on trees, branch lengths and model parameters, to give the **posterior** probability of each tree:

$$P(T, \theta \mid D) = \frac{P(D \mid T, \theta)\, P(T, \theta)}{P(D)}.$$

The denominator, a sum over all trees and an integral over all parameters, cannot be computed. Instead, trees are sampled from the posterior with **Markov chain Monte Carlo** (MCMC). The chain proposes a small change, for example a new branch length or a rearrangement of the topology, and accepts it with a probability given by the Metropolis–Hastings rule, which depends only on the ratio of the posterior densities. After a *burn-in* period the chain visits trees in proportion to their posterior probability, so the posterior probability of a clade is simply the fraction of sampled trees that contain it (Rannala and Yang 1996; Huelsenbeck et al. 2001).

The main program for general phylogenetics is **MrBayes** (Ronquist et al. 2012). **BEAST** and BEAST 2 (Suchard et al. 2018; Bouckaert et al. 2019) specialise in time-calibrated trees: they combine relaxed molecular clocks, in which the rate may vary between branches (Drummond et al. 2006), fossil or sampling-date calibrations, and population models such as the coalescent. BEAST has become the standard tool for *phylodynamics*, the reconstruction of the spread of epidemics such as influenza, Ebola and SARS-CoV-2 from dated virus sequences. PhyloBayes implements site-heterogeneous models such as CAT (Lartillot and Philippe 2004) that are important for deep phylogenies (Section 4.4).

Bayesian methods have three main advantages: they give a direct probabilistic measure of uncertainty, they handle complex, parameter-rich models by integrating over parameters rather than optimising them, and they allow prior knowledge, such as fossil dates, to be included. The costs are the need to choose priors, which can influence the result, and the need to check that the MCMC chains have **converged**: several independent chains should be run, and their samples compared (for example with the average standard deviation of split frequencies, or effective sample sizes).

4.3.5 Comparing the methods

Table 4.1: Main families of tree-building methods.

Method	Criterion	Model	Strengths	Weaknesses
UPGMA	clustering of distances	clock assumed	very fast	wrong when rates vary
Neighbour joining	minimum evolution (greedy)	for distance correction	fast, consistent for additive distances	information lost in distances
Maximum parsimony	fewest changes	none	intuitive, no model needed	long-branch attraction, inconsistent
Maximum likelihood	probability of the data	explicit	uses all data, consistent, testable	slow, model-dependent
Bayesian inference	posterior probability	explicit, with priors	uncertainty, complex models, dating	slow, priors, convergence

See also: further explanation in Section D.4.2; related Wikipedia articles in Section E.4.2.

i Sources

Parts of this section are adapted from the Wikipedia articles Computational phylogenetics, Distance matrices in phylogeny, UPGMA, Neighbor joining, Maximum parsimony, Long branch attraction, Felsenstein's tree-pruning algorithm and Bayesian inference in phylogeny (CC BY-SA 4.0, retrieved 3 October 2026).

4.3.6 Substitution models

Distance corrections, maximum likelihood and Bayesian inference all rely on a **substitution model**: a description of how one nucleotide or amino acid is replaced by another over time. Almost all models used in practice are continuous-time Markov processes. They are defined by a rate matrix Q, whose element q_{xy} is the instantaneous rate of change from state x to state y. The probability of change along a branch of length t is then

$$P(t) = e^{Qt},$$

computed by diagonalising Q. The models are normally assumed to be *stationary* (the base composition does not change over the tree), *time-reversible* ($\pi_x q_{xy} = \pi_y q_{yx}$, which is why the root position does not affect the likelihood) and the same at all sites, except for the rate. The main nucleotide models form a nested series:

Table 4.2: Nested nucleotide substitution models. JC69 = Jukes–Cantor, K80 = Kimura two-parameter, HKY = Hasegawa–Kishino–Yano, GTR = general time-reversible.

Model	Base frequencies	Exchange rates	Parameters
JC69 (Jukes and Cantor 1969)	equal	all equal	0

Model	Base frequencies	Exchange rates	Parameters
K80 (Kimura 1980)	equal	transitions ≠ transversions	1
F81 (Felsenstein 1981)	unequal	all equal	3
HKY85 (Hasegawa et al. 1985)	unequal	transitions ≠ transversions	4
GTR (Tavaré 1986)	unequal	all six different	8

Real sequences also differ in rate from site to site: some positions are invariant because of strong functional constraints, while others, such as third codon positions, change rapidly (Section 3.5). This **rate heterogeneity** is usually modelled with a **gamma distribution** of rates across sites (Yang 1994), approximated by four to eight discrete rate categories, sometimes combined with a proportion of invariable sites. The models are named by combining the parts, for example GTR+Γ or GTR+I+Γ. Ignoring rate heterogeneity makes distances too short for divergent sequences and is a common cause of long-branch attraction.

For **proteins**, the 190 exchange rates between amino acids are too many to estimate from a single alignment. Instead, *empirical* matrices are estimated from large databases of alignments, in the tradition of Dayhoff's PAM matrices (Section 1.3): JTT, WAG (Whelan and Goldman 2001) and LG (Le and Gascuel 2008) are the most common. Specialised matrices exist for mitochondrial, chloroplast and viral proteins. Codon models, which describe synonymous and non-synonymous changes separately, are used to detect selection (Section 3.6).

The assumption that all sites follow the same process is clearly false. In particular, different sites in a protein accept different sets of amino acids. **Site-heterogeneous mixture models**, such as CAT (Lartillot and Philippe 2004) and the C10–C60 profile mixtures, allow each site to belong to one of many classes with its own amino acid preferences. They fit real data much better and are less sensitive to long-branch attraction in deep phylogenies (Section 4.4). Other extensions allow the rate of a site to change over the tree (*heterotachy*), or the composition to differ between lineages (non-stationary models).

Model selection. A more complex model never fits the data worse than a simpler model nested within it, but at the cost of more parameters. Nested models can be compared with the likelihood-ratio test. More generally, models are compared with the Akaike information criterion, $\mathrm{AIC} = 2k - 2\ln L$, or the Bayesian information criterion, $\mathrm{BIC} = k\ln m - 2\ln L$, where k is the number of free parameters and m the number of sites. ModelTest (Posada and Crandall 1998) made this routine, and ModelFinder (Kalyaanamoorthy et al. 2017), in IQ-TREE, also tests rate-heterogeneity models and different models for different partitions (for example genes or codon positions) of the alignment. In Bayesian analyses, models can be compared with Bayes factors or averaged over with reversible-jump MCMC.

Model adequacy is a separate question. Model selection only finds the best of the candidate models; it does not show that this model describes the data well, and the best of several poor models is still poor. Whether a model is adequate is tested by asking if it could have produced the observed data: data sets are simulated under the fitted model (a *parametric bootstrap*, or *posterior predictive* simulation in a Bayesian analysis) and compared with the real alignment, for example in base composition across taxa or in the amount of saturation (Brown 2014). Inadequate models are a main source of the systematic errors described in Section 4.4.

4.3.7 DNA, codons or proteins?

Protein-coding genes can be analysed as nucleotides, as codons or as amino acids, and the choice matters as much as the model.

- **Nucleotides** give three characters per codon and record synonymous changes, so they hold the most information for closely related sequences: populations, species within a genus, or viral outbreaks. Their problem is **saturation**. With only four states, and with third codon positions changing rapidly, sites that have changed several times look no more different than sites that have changed once, and the similarity that remains is increasingly due to chance and to shared composition. Saturation can be seen by plotting the observed (uncorrected) distances against model-corrected distances: the curve flattens when the sequences are saturated.
- **Amino acids** have 20 states, ignore synonymous changes and evolve more slowly, so they saturate much later. They are the usual choice for deep divergences, from hundreds of millions of years onwards, and with site-heterogeneous mixture models for the deepest questions (Section 4.5). The price is that information is lost for close relationships.
- **Codon models** describe substitutions between the 61 sense codons, using the genetic code, the transition/transversion ratio and the ratio $\omega = d_N/d_S$ of non-synonymous to synonymous rates (Goldman and Yang 1994; "A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome." 1994). They are the most realistic description of coding sequences, and are essential for detecting selection (Section 3.6), but they are much slower to compute and are less often used for tree search itself.

Compositional heterogeneity is a second problem. Standard models assume one base or amino acid composition for the whole tree (stationarity), but genomes differ widely in GC content, and the composition of proteins follows it partly. Unrelated lineages with similar composition, for example GC-rich thermophiles, are then pulled together. Remedies include removing or recoding third codon positions (RY-coding, in which only purines and pyrimidines are distinguished), LogDet distances, non-stationary models that allow the composition to change over the tree (Foster 2004), and recoding amino acids into a few groups of similar residues. Partitioning the data, with separate models for first, second and third codon positions, is a common compromise for nucleotide data of intermediate depth. A simple rule is to use nucleotides when the sequences are close, codon positions or codon models at intermediate distances, and proteins when third positions are saturated, and to check the conclusion with more than one data type.

4.3.8 Testing the molecular clock

Many analyses, such as UPGMA trees, clock rooting and divergence dating, assume a **molecular clock** (Section 3.3), and this assumption can be tested.

The **relative rate test** (Sarich and Wilson 1973) needs only three sequences: two ingroup sequences A and B, and an outgroup C. Under a clock, A and B have changed equally since they diverged, so the distances d_{AC} and d_{BC} should be equal, within their sampling error. Tajima's version counts the sites where only A or only B differs from the other two sequences, and compares the two counts with a χ^2 test (Tajima 1993). The test does not require a tree or a model of the dates.

For a whole tree, a **likelihood-ratio test** compares the maximum likelihood of the tree without a clock, which has $2n - 3$ free branch lengths for n sequences, with that of the same topology constrained to be ultrametric, which has $n - 1$ node heights. The clock model is nested in the non-clock model, so twice the difference in log-likelihood is compared with a χ^2 distribution with $n - 2$ degrees of freedom. For time-stamped samples, such as viruses collected over several years, a regression of root-to-tip distance against sampling date shows whether there is enough clock-like signal to date the tree.

With large data sets the strict clock is almost always rejected, as the first tests showed (Langley and Fitch 1974). Rejection does not mean that dating is impossible, but that rates must be modelled: by *local clocks* for different parts of the tree, or by the *relaxed clocks* used in Bayesian dating (Section 4.3), with the dates then depending more heavily on fossil calibrations and on the clock model.

4.3.9 Statistical support

A tree estimated from a finite alignment is uncertain, and different branches are uncertain to different degrees. Several measures are used to express this.

- **The nonparametric bootstrap** (Felsenstein 1985) creates many pseudo-replicate alignments by sampling columns from the original alignment with replacement, builds a tree from each, and reports for each branch the percentage of replicate trees that contain it. Values above about 70% have traditionally been taken as good support. The bootstrap measures how stable a branch is when the sites are resampled, for a given alignment, model and tree-building method. It is *not* the probability that the branch is correct: a wrong model or alignment can give a wrong branch 100% bootstrap support. In simulations it tends to be conservative for well-supported branches. With hundreds of replicates it is expensive for ML, which led to faster approximations such as the **ultrafast bootstrap** of IQ-TREE (Hoang et al. 2017). Its values are on a different scale and are usually trusted only from about 95%, compared with about 70% for the standard bootstrap. For very large trees, where a single misplaced taxon destroys a branch in all replicates, the **transfer bootstrap** counts how many taxa must be moved to recover the branch (Lemoine et al. 2018).
- **Approximate likelihood-ratio tests** (aLRT, SH-aLRT) compare, for each internal branch, the likelihood of the best tree with that of the best alternative arrangement around that branch (Anisimova and Gascuel 2006; Guindon et al. 2010). They are fast and widely reported together with the ultrafast bootstrap.
- **Posterior probabilities** from Bayesian analyses are the probability that a clade is correct *given the model and priors*. They are easier to interpret than bootstrap values, but they tend to be high (often 1.0) for short branches with little support, especially when the model is misspecified. Bootstrap values and posterior probabilities are therefore not directly comparable.
- **Topology tests** compare whole trees, for example the ML tree with a tree that groups taxa as an alternative hypothesis predicts. The Kishino–Hasegawa (KH), Shimodaira–Hasegawa (SH) and approximately unbiased (AU) tests (Shimodaira 2002) use the per-site likelihoods of the trees, and ask whether the difference is larger than expected by chance.

Support values and branch lengths are different quantities that are easy to confuse on a printed tree. The branch length is drawn as the length of the branch, with a scale bar in substitutions per site. The support value is a number written next to a node or branch, and in Newick files it is stored as the label of the internal node, as in `(A:0.1,B:0.2)95:0.05`. A long branch can have low support and a very short branch high support.

All of these measures assess *random error*, which decreases as more data are added. None of them protects against *systematic error* from a wrong model or a wrong alignment, which in phylogenomics can give 100% support for an incorrect tree (Section 4.4).

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Models of DNA evolution, Substitution model, Model selection and Bootstrapping (statistics) (CC BY-SA 4.0, retrieved 3 October 2026). The subsections on data types and on testing the molecular clock were written for this book from the primary literature cited in the text.

4.4 Alignments, genomes and conflicting trees

The classical methods were developed for one gene and tens of taxa. Today, trees are routinely built from hundreds or thousands of genes. With so much data, random error is no longer the main problem.

Errors in the input alignments, model misspecification and genuine conflict between the histories of different genes become more important (Delsuc et al. 2005; Jeffroy et al. 2006; Philippe et al. 2011; Kapli et al. 2020).

4.4.1 Alignment uncertainty

Phylogenetic methods treat the multiple alignment as known, but the alignment is itself an estimate (Section 1.6). In variable regions, different programs, or even different guide trees in the same program, give different alignments, and the trees built from them can differ (Wong et al. 2008). Aligners also interact with the tree in non-obvious ways. Progressive aligners that do not distinguish insertions from deletions tend to "over-align" gaps, and PRANK was designed with a phylogeny-aware gap treatment to avoid this (Löytynoja and Goldman 2008).

There are three ways of dealing with alignment uncertainty:

- **Filtering**: removing poorly aligned and highly variable columns before building the tree, with programs such as Gblocks (Talavera and Castresana 2007) or trimAl (Capella-Gutiérrez et al. 2009). Moderate trimming usually helps, but aggressive trimming removes real phylogenetic signal and can make trees worse.
- **Comparing alternatives**: building trees from several alignments and reporting only the conclusions that are robust to the choice.
- **Joint estimation**: treating the alignment as a parameter and co-estimating alignment and tree in a statistical framework, starting with the TKF91 model of insertions and deletions (Thorne et al. 1991). BAli-Phy samples alignments and trees jointly with MCMC (Redelings and Suchard 2005). This is the most principled solution, but it is computationally feasible only for small data sets.

4.4.2 Phylogenomics

Phylogenomics uses genome-scale data to infer phylogenies. The first step is to identify sets of orthologous genes, usually single-copy genes present in most of the species (Section 3.4), because paralogues that are mistaken for orthologues (*hidden paralogy*) give wrong trees. There are two main ways to combine the genes:

- **Concatenation** (the *supermatrix* approach): the alignments of all genes are joined end to end and analysed as one, often with a separate model for each gene (a *partitioned* analysis). It assumes that all genes share one tree.
- **Separate analysis**: a tree is built for each gene, and the gene trees are combined into a species tree, either as a *supertree* or with methods based on the multispecies coalescent (below).

Early phylogenomic studies showed that single genes are often in conflict, but that concatenating a few dozen genes gave a single, well-supported tree (Rokas et al. 2003). It was soon found that concatenation can give 100% support to conflicting trees in different studies (Jeffroy et al. 2006). With many genes, **systematic errors** dominate, because they do not decrease with more data (Philippe et al. 2011):

- **long-branch attraction** (Section 4.3), when some lineages evolve much faster than others;
- **compositional heterogeneity**, when unrelated lineages share a similar GC or amino acid composition and are grouped together;
- **heterotachy**, when the rate of a site changes over time;
- **saturation**, when sites have changed so often that the signal of old divergences is lost.

The remedies are better models (site-heterogeneous mixtures, non-stationary models; Section 4.3.6), denser taxon sampling to break up long branches, removal of the fastest-evolving sites or genes, and recoding amino acids into groups to reduce compositional bias. The position of animals and fungi, the root of the animal tree (sponges or comb jellies first?) and the relationships between eukaryotic supergroups are well-known examples where the answer depends on how these errors are handled (Steenwyk et al. 2023).

4.4.3 Gene trees and species trees

Even with perfect data and a perfect model, a gene tree need not have the same shape as the species tree (Maddison 1997; Degnan and Rosenberg 2009). Several distinct biological processes cause conflict, and they leave different signatures:

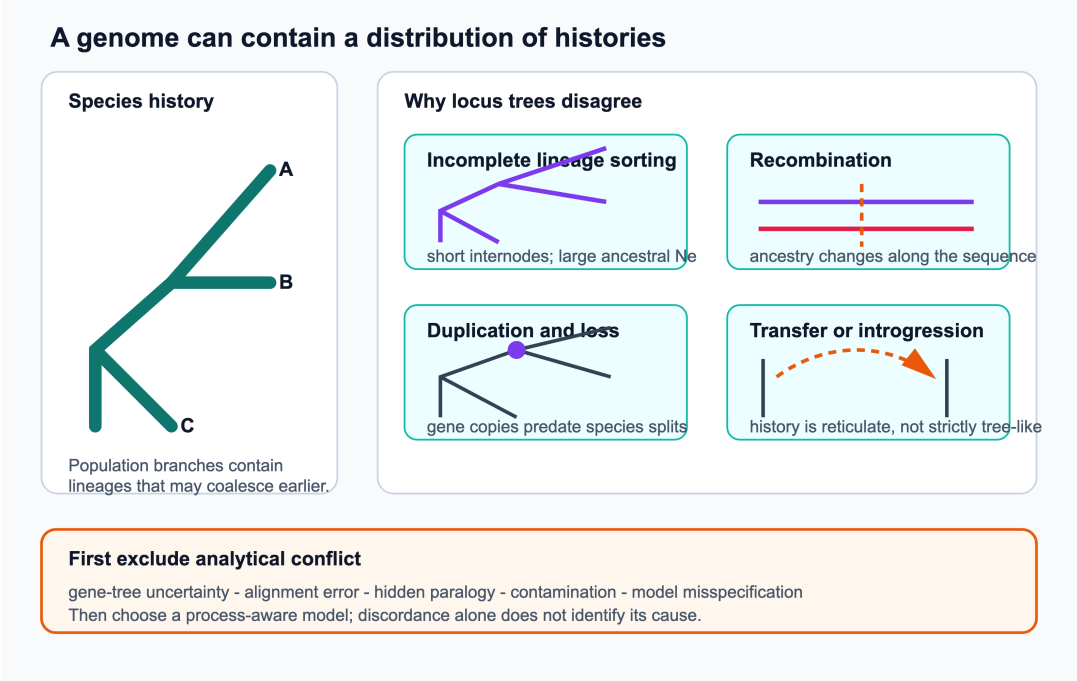


Figure 4.2: Why gene trees disagree with the species tree. The branches of a species tree are populations, in which the lineages of a gene can coalesce earlier than the speciations; incomplete lineage sorting, recombination, duplication and loss, and transfer or introgression all produce gene trees that differ from the species tree. Gene-tree error, alignment error, hidden paralogy, contamination and model misspecification must be excluded first. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

- **Incomplete lineage sorting** (ILS). A gene tree traces the ancestry of gene copies back in time, and copies coalesce into their common ancestor at some time *before* the speciation that separated their carriers. If two speciation events follow each other quickly, compared with the effective population size, ancestral polymorphisms can survive through both speciations and be sorted randomly into the descendant species. A gene can then group humans with gorillas rather than with chimpanzees, as about a third of the human genome does (Section 4.5). The **multispecies coalescent** model of population genetics predicts how often each gene tree occurs. For three species, if the internal branch of the species tree has length t in coalescent units (generations divided by $2N_e$), the gene tree that matches the species tree has probability $1 - \frac{2}{3}e^{-t}$ and each of the two discordant trees $\frac{1}{3}e^{-t}$, so short internal branches and large ancestral populations give much discordance. In extreme cases the most common gene tree differs from the species tree

(the *anomaly zone*) (Degnan and Rosenberg 2006), and concatenation can then be statistically inconsistent.

- **Gene duplication and loss.** If a gene duplicated before a speciation and different copies were lost in different lineages, the surviving genes are paralogues whose tree reflects the duplication, not the speciation (hidden paralogy; Section 3.4).
- **Horizontal gene transfer** (HGT), common in prokaryotes, places a gene in a lineage it did not descend in (Section 3.7). Hybridisation and introgression have the same effect in eukaryotes.
- **Recombination** within a locus means that different parts of one gene have different histories, so that no single tree describes the whole gene. Methods that assume one tree per gene, including ASTRAL below, therefore assume that each locus is free of internal recombination.

In practice, much of the observed conflict is not biological at all but **gene-tree estimation error**: single genes are short, carry little signal, and are affected by the same alignment and model errors as any other tree. The causes can often be told apart. ILS is concentrated around short internal branches, and for three species it makes the two minority gene trees about equally frequent; introgression makes one of them more frequent than the other (the basis of the ABBA–BABA test (Green et al. 2010; Durand et al. 2011)). Duplication and loss and HGT are tied to specific events and lineages. Estimation error gives conflict that is poorly supported and changes with the data or model. These causes should be distinguished before HGT, or any other single process, is invoked to explain discordance.

Species-tree methods estimate the species tree from many gene trees while accounting for these processes (Edwards 2009). The most widely used are *summary methods* based on the multispecies coalescent, such as ASTRAL (Mirarab et al. 2014; Zhang et al. 2018), which finds the species tree that agrees with the largest number of four-taxon subtrees (quartets) of the gene trees. Because the most frequent unrooted gene tree for any four taxa always matches the species tree under the coalescent, this is consistent under ILS. Full Bayesian methods that co-estimate gene trees and species tree (such as *BEAST) are more accurate but much slower. Concatenation and coalescent methods give the same answer for most well-resolved relationships, but they can differ for rapid radiations.

4.4.4 Consensus trees, supertrees and concordance

When many trees are available, from bootstrap replicates, an MCMC sample or many genes, they have to be summarised. A **consensus tree** contains the splits that occur in a given fraction of the input trees: all of them in the *strict* consensus, more than half in the *majority-rule* consensus (Margush and McMorris 1981). **Supertree** methods combine trees with different but overlapping sets of taxa into one tree (Bininda-Emonds 2004). Like the coalescent methods above, they treat the estimated gene trees as the data, and are sensitive to poorly estimated gene trees.

The numbers on such trees must be read with care. On a consensus of gene trees, the percentage at a node is the fraction of *genes* that support a split, not a bootstrap value, which measures how stable the split is when the *sites* of one alignment are resampled. Resampling whole genes and resampling sites ask different questions. In a large concatenated data set, almost every branch gets 100% bootstrap support, even when many genes disagree. **Concordance factors** describe the conflict instead: the *gene concordance factor* of a branch is the percentage of gene trees that contain it, and the *site concordance factor* the percentage of informative sites that support it rather than the two alternative arrangements (Minh, Hahn, et al. 2020). A branch with 100% bootstrap support but a gene concordance factor of 35% is well supported by the data as a whole but contested by most individual genes, a typical sign of a rapid radiation or of systematic error.

4.4.5 Reconciliation

Reconciliation compares a gene tree with a species tree and infers the events that explain the differences. Goodman and colleagues introduced the idea in 1979, using globins (Goodman et al. 1979). Each node of the gene tree is mapped to the species tree: if a gene tree node's two descendant subtrees contain genes from overlapping sets of species, the node is interpreted as a duplication, otherwise as a speciation. Modern methods infer duplications and losses (DL models), or duplications, transfers and losses (DTL models), by finding the most parsimonious or most likely set of events. Reconciliation is the basis of tree-based orthology inference (Section 3.4). It also works in the other direction: errors in gene trees can be corrected by using the species tree as a prior, as in methods that infer gene trees together with the species tree (Szöllősi et al. 2014).

4.4.6 Phylogenomics for function prediction

The word phylogenomics was first used for a different idea: the use of gene trees to predict function (Eisen 1998). Annotation by the best database hit (Chapter 10) ignores the history of the family. If a family contains two subfamilies created by an old duplication, with different functions, an uncharacterised protein that falls in subfamily B should not get the function of a protein from subfamily A, even if that protein happens to be its best hit. The phylogenomic approach builds a tree of the family, marks the members with experimentally known functions, infers duplication and speciation nodes by reconciliation, and transfers a function only within the clade, usually a group of orthologues, in which it is supported. Probabilistic versions such as SIFTER model how function changes along the tree (Engelhardt et al. 2005), and the Gene Ontology consortium propagates annotations manually along PANTHER family trees (Gaudet et al. 2011). The method depends on the quality of the tree, the taxon sampling and the rooting, and a function can also change within a group of orthologues.

4.4.7 Phylogenetic networks

When the history of a set of sequences includes recombination, hybridisation or HGT, a single tree cannot represent it. **Phylogenetic networks** generalise trees in two ways (Huson and Bryant 2005):

- **Split networks**, built for example with Neighbor-Net (Bryant 2003) and SplitsTree, display conflicting signals in the data as boxes, without claiming that each box corresponds to a specific evolutionary event. They are useful as exploratory tools: a tree-like network shows that the data are consistent with a tree, while a tangled one shows strong conflict.
- **Explicit networks**, such as hybridisation, recombination and ancestral recombination graphs, represent specific reticulation events with nodes that have two parents.

Recombination is a particular problem for viruses and for bacteria that exchange DNA frequently. A tree built from a recombinant alignment is a mixture of several histories and can be wrong even with strong support. The usual practice is to detect recombination breakpoints first (Kosakovsky Pond et al. 2006), and then to build separate trees for non-recombining segments, to remove the recombined regions, or to use a network.

For prokaryotes, the question of whether evolution is best described by a tree or a network is central to the debate about the tree of life (W. F. Doolittle 1999); see Section 3.7 and Section 4.5.

See also: further explanation in Section D.4.3; related Wikipedia articles in Section E.4.3.

Sources

Parts of this section are adapted from the Wikipedia articles Phylogenomics, Multiple sequence alignment, Incomplete lineage sorting, Multispecies coalescent process, Phylogenetic reconciliation and Phylogenetic network (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Phylogenomics and Discordant Histories" (CC BY 4.0), used with permission.

4.5 The tree of life

Molecular phylogenetics has changed our view of the relationships between all living organisms. This section follows the main steps, from the discovery of the archaea to the current debates about the root of the tree and the origin of eukaryotes, and ends with the best-studied small tree of all, that of humans and the other great apes.

4.5.1 Three domains

Until the 1970s, life was divided into prokaryotes and eukaryotes, and the prokaryotes could not be classified phylogenetically at all, since they have few informative morphological characters. In 1977 Carl Woese and George Fox compared fragments of the 16S ribosomal RNA from many prokaryotes and found that the methanogens were as different from other bacteria as from eukaryotes (Woese and Fox 1977). Ribosomal RNA was an ideal marker: it is present in all cellular life, it is essential and rarely transferred, and it has both conserved and variable regions. Woese proposed the **three-domain system** of Bacteria, Archaea and Eukarya (Woese et al. 1990), which replaced the prokaryote–eukaryote dichotomy and the five-kingdom system.

Since 16S rRNA genes can be amplified directly from the environment, the rRNA tree also opened up the study of uncultivated microbes. Most microbial diversity turned out to be invisible to cultivation. Metagenomics later added whole genomes from uncultivated organisms: a tree built from ribosomal proteins of over 3,000 genomes showed that most diversity is bacterial, and that a large fraction belongs to the **candidate phyla radiation** of small, mostly symbiotic bacteria (Brown et al. 2015; Hug et al. 2016). Genome-based trees are now used to define prokaryotic taxonomy itself, as in the Genome Taxonomy Database (GTDB) (Parks et al. 2018).

4.5.2 Rooting the tree of life

A tree of all life has no outgroup, so it cannot be rooted in the usual way. In 1989 two groups addressed the problem with **paralogous genes** that duplicated before the last universal common ancestor (Gogarten et al. 1989; Iwabe et al. 1989). Each copy of such a pair, for example the two elongation factors EF-Tu and EF-G, or the catalytic and regulatory subunits of ATP synthase, can serve as an outgroup for the tree of the other. Both studies placed the root between the Bacteria and the Archaea + Eukarya, which has been the standard view since, although the signal in such ancient duplications is weak and the analyses are sensitive to long-branch attraction.

The organism at the root, the **last universal common ancestor (LUCA)**, can be characterised by the genes that are reconstructed as present at the root. Weiss and colleagues identified 355 protein families that suggest an anaerobic, thermophilic organism that fixed carbon and nitrogen and depended on hydrogen, possibly in a hydrothermal environment (Weiss et al. 2016). A more recent analysis using gene-tree–species-tree reconciliation estimated that LUCA lived about 4.2 billion years ago and already

had a relatively complex genome, including an early immune system (Moody et al. 2024). Both the age and the gene content are uncertain, because HGT and gene loss make it hard to decide which genes really go back to LUCA. The root of the Bacteria themselves has been placed with similar methods between two large groups, the Terrabacteria and the Gracilicutes (Coleman et al. 2021). The origin of life before LUCA is discussed in Section 3.2.

4.5.3 Two domains and the origin of eukaryotes

Where do eukaryotes come from? The three-domain tree makes Archaea and Eukarya sister groups. The alternative **eocyte hypothesis**, proposed by James Lake in the 1980s, places eukaryotes *within* the Archaea. With better models and more genomes, phylogenomic analyses have come to support the second view, the **two-domain tree** (Williams et al. 2019).

The decisive data came from metagenomics. In 2015, genomes assembled from deep-sea sediments near the hydrothermal vent field Loki's Castle revealed the **Lokiarchaeota**, archaea that encode many proteins previously thought to be specific to eukaryotes, such as actins and small GTPases (Spang et al. 2015). The related **Asgard archaea** were then found worldwide (Zaremba-Niedzwiedzka et al. 2017), the first one was cultivated after a twelve-year effort (Imachi et al. 2020), and the most recent analyses place eukaryotes within the Asgard archaea, from within the Heimdallarchaeia (Eme et al. 2023). In this view, eukaryotes arose when an Asgard archaeon took up an alphaproteobacterium that became the mitochondrion. Eukaryotic genomes are therefore chimeras: genes for information processing (replication, transcription, translation) are mostly archaeal, while many metabolic genes are bacterial, partly through *endosymbiotic gene transfer* (Section 3.7).

Within the eukaryotes, the classical kingdoms (animals, plants, fungi, protists) have been replaced by a handful of **supergroups**. Animals and fungi belong to the Opisthokonta, plants to the Archaeplastida, and most of the diversity is in protist lineages such as the SAR group (stramenopiles, alveolates and rhizarians) and several lineages once grouped together as the Excavata. The relationships between the supergroups, and the position of the eukaryotic root, are still debated (Burki et al. 2020). Combining molecular clocks with fossils, the last common ancestor of eukaryotes is estimated to have lived roughly 1–2 billion years ago, although the estimates vary widely (Betts et al. 2018).

4.5.4 Tree or network?

The tree of life is a tree only for some genes. In prokaryotes, frequent horizontal gene transfer means that different genes have different histories, and some authors have argued that the tree of life should be replaced by a network (W. F. Doolittle 1999). Most phylogenomic analyses take a middle position: a core of genes that are rarely transferred, such as those for ribosomal proteins, gives a consistent "tree of cells", on top of which there is a web of transfers (Section 3.7; Section 4.4). The origin of eukaryotes, by fusion of an archaeon and a bacterium, is itself a reticulation at the base of the eukaryotic branch.

4.5.5 Taxonomy and the tree

Taxonomy is the science of naming, describing and classifying organisms. Phylogeny and taxonomy are related but not the same: a phylogeny is a hypothesis about history, while a taxonomy is a system of names, governed by conventions, that is used to communicate about organisms. Since Hennig, most taxonomists aim for a classification in which every named group is a clade (Section 4.2), so taxonomies are revised as trees change. Paraphyletic groups such as "reptiles" without birds, or the "Pongidae" discussed below, are then abandoned or kept only as informal names.

The **Linnaean ranks** (domain, kingdom, phylum, class, order, family, genus, species) are a legacy of pre-evolutionary classification. A tree has no ranks, only nested clades, and most clades are left unranked. Ranks are assigned by convention within each group, so they are **not comparable across groups**: a mammalian genus and a bacterial genus need not be of similar age or divergence, and a "family" of insects may be much older than a "family" of birds. Counting families or genera to compare diversity between groups is therefore misleading.

Names are governed by separate **codes of nomenclature**: for animals (ICZN), for algae, fungi and plants (ICN), for prokaryotes (ICNP) and for viruses (ICTV). The codes regulate names, priority and type specimens, but not the classification itself. The prokaryotic code requires a described type strain deposited in culture collections, which excluded most of the diversity known only from sequences; such organisms received provisional *Candidatus* names. The **SeqCode** now allows a genome sequence to serve as the type, so that uncultivated prokaryotes can be validly named (Hedlund et al. 2022).

Species are the most debated rank. The biological species concept, based on reproductive isolation, cannot be applied to asexual organisms or to most prokaryotes, which exchange genes horizontally across species boundaries (Section 3.7). Prokaryotic species were traditionally defined operationally, by 70% DNA–DNA hybridisation and later also by 16S rRNA identity (about 97–99%; Section 5.7). Genome sequences replaced these with the **average nucleotide identity** (ANI, Section 1.5) of shared regions: an ANI of about 95% corresponds to the old hybridisation threshold (Konstantinidis and Tiedje 2005), and comparisons of tens of thousands of genomes show a gap in the distribution of ANI values around 95%, suggesting a natural, if fuzzy, species boundary (C. Jain et al. 2018). Whether this gap reflects biology or the uneven sampling of genomes is still discussed.

The **Genome Taxonomy Database** (GTDB) applies these ideas systematically to bacteria and archaea (Parks et al. 2018; Parks et al. 2020). It builds trees from conserved single-copy proteins (120 for bacteria, 53 for archaea) of all available genomes, including metagenome-assembled ones, and requires every taxon to be monophyletic. Ranks are normalised with the **relative evolutionary divergence** (RED): each node in the tree is given a value between 0 at the root and 1 at the tips, interpolated along the branches so that lineages evolving at different rates are treated fairly, and taxa of the same rank are required to have similar RED values. Species are defined as clusters of genomes with ANI of at least about 95%. The result changed the classification of a large fraction of genomes, split polyphyletic groups (often marked with suffixes, such as Firmicutes_A) and merged others. It is the closest thing to a phylogenetically consistent taxonomy for any large group of organisms, but it differs from NCBI Taxonomy, which is a curated working classification and states that it is not an authority on names (Section 2.4).

Sequence-based taxonomy has changed what is classified. After Woese, 16S rRNA surveys showed that most microbial lineages had never been cultivated, and metagenomics now provides genomes for them (Section 5.7). A large part of the named phyla of bacteria and archaea is known only from sequences, and the same is true for viruses, many of which are now classified from metagenomic genomes alone. In practice, a taxonomic label in a metagenomic study is an inference that depends on the reference taxonomy (NCBI or GTDB) and its version, and it should be reported as such.

4.5.6 Hominoid and human evolution

The great apes (family Hominidae) are humans, chimpanzees and bonobos, gorillas and orangutans. Together with the gibbons they form the hominoids. Until the 1960s, humans were usually placed in their own family, and the other great apes in the Pongidae. This turned out to be a paraphyletic group: immunological and protein data from Morris Goodman and others, and later DNA sequences, showed that chimpanzees are more closely related to humans than to gorillas. The current consensus tree is ((((human, (chimpanzee, bonobo)), gorilla), orangutan), gibbons). Classical estimates put the divergence of humans and chimpanzees at about 5–7 million years ago, gorillas at 8–10 million years

ago and orangutans at 15–20 million years ago, but these dates depend strongly on the mutation rate and generation time used (Scally and Durbin 2012).

The genomes of the chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005), gorilla (Scally et al. 2012) and other great apes (Prado-Martinez et al. 2013) made the great-ape tree a textbook case of **incomplete lineage sorting** (Section 4.4). Because the human–chimpanzee and human–gorilla speciations were close in time and ancestral populations were large, about 30% of the gorilla genome is closer to human or chimpanzee than these are to each other (Scally et al. 2012). Models that combine the coalescent with a hidden Markov model along the genome (CoalHMM) estimate speciation times and ancestral population sizes from these patterns (Hobolth et al. 2007). Patterns on the X chromosome even suggested that gene flow continued between early human and chimpanzee populations after their initial split (Patterson et al. 2006), a claim that is still debated. The first complete, telomere-to-telomere genomes of all great apes now give access also to the repetitive and structurally variable parts of the genomes (Yoo et al. 2025).

The fossil record of the *hominins*, the lineage leading to humans after the split from chimpanzees, includes early forms such as *Sahelanthropus* and *Ardipithecus*, several species of *Australopithecus* (including "Lucy", *A. afarensis*, 3.2 million years old) and the genus *Homo*, which appeared about 2.5 million years ago. Several *Homo* species coexisted until recently. Anatomically modern humans arose in Africa about 300,000 years ago and spread across the world from about 60,000 years ago. Ancient DNA has shown that they met and mixed with Neanderthals and Denisovans. The genomics of these events, and what they reveal about human genome evolution, is discussed in Chapter 6.

See also: further explanation in Section D.4.4; related Wikipedia articles in Section E.4.4.

i Sources

Parts of this section are adapted from the Wikipedia articles Three-domain system, Two-domain system, Tree of life (biology), Last universal common ancestor, Asgard (archaea), Eukaryote, Hominidae and Human evolution (CC BY-SA 4.0, retrieved 3 October 2026). The subsection on taxonomy was written for this book from the primary literature cited in the text.

4.6 Key concepts

A checklist for revision.

- **History:** Haeckel's trees; Hennig and cladistics; Fitch and Margoliash (1967); Felsenstein's contributions (long-branch attraction, pruning algorithm, bootstrap); neighbour joining; Bayesian phylogenetics
- **Tree terminology:** rooted/unrooted, topology, branch lengths (change, not time) and ultrametric time trees; clades, mono-, para- and polyphyly; splits vs. clades; sister groups; polytomies; outgroup, midpoint and clock rooting and their assumptions; the number of possible trees, $(2n-5)!!$; Newick format
- **Distance methods:** corrected distances and saturation (Jukes–Cantor correction); UPGMA and the molecular-clock (ultrametric) assumption; neighbour joining, additive distances and its Q criterion; minimum evolution and FastME
- **Maximum parsimony:** Fitch's and Sankoff's algorithms; informative sites; long-branch attraction and the Felsenstein zone; statistical inconsistency
- **Maximum likelihood:** the likelihood of a tree; Felsenstein's pruning algorithm; tree search (NNI, SPR, TBR); RAxML, IQ-TREE, PhyML, FastTree
- **Bayesian inference:** priors and posteriors; MCMC, burn-in and convergence; MrBayes; dating with relaxed clocks in BEAST; phylodynamics

- **Substitution models:** Markov models, $P(t) = e^{Qt}$, stationarity and reversibility; JC69, K80, F81, HKY, GTR; rate heterogeneity (+Γ, +I); empirical protein matrices (JTT, WAG, LG); site-heterogeneous mixture models (CAT); model selection with LRT, AIC and BIC; model adequacy (parametric bootstrap, posterior predictive checks)
- **Data types and clocks:** nucleotides vs. codons vs. amino acids and when each is appropriate; saturation and how to detect it; compositional heterogeneity and its remedies (removing third positions, RY-coding, LogDet, non-stationary models, recoding); partitioning by codon position; codon models and ω; relative rate tests (Sarich–Wilson, Tajima); likelihood-ratio test of clock vs. no clock ($n - 2$ degrees of freedom); root-to-tip regression; local and relaxed clocks
- **Support:** nonparametric and ultrafast bootstrap (not the probability that a clade is true); support values vs. branch lengths; transfer bootstrap; aLRT; posterior probabilities and why they differ from bootstrap values; topology tests (KH, SH, AU); random vs. systematic error
- **Alignment uncertainty:** alignment errors and trees; trimming (Gblocks, trimAl); joint estimation of alignment and tree (BAli-Phy)
- **Phylogenomics:** orthologue selection and hidden paralogy; concatenation vs. separate analysis; systematic errors (long-branch attraction, compositional bias, heterotachy, saturation) and their remedies
- **Gene trees vs. species trees:** incomplete lineage sorting and the multispecies coalescent; duplication and loss; HGT and introgression; recombination; three-taxon gene-tree probabilities under the coalescent; gene-tree estimation error; ASTRAL; strict and majority-rule consensus trees and supertrees; gene versus site resampling; gene and site concordance factors; reconciliation (DL and DTL models); split and explicit phylogenetic networks
- **Phylogenomic function prediction:** Eisen's phylogenomics; transferring function within a clade of orthologues rather than from the best hit; SIFTER; GO annotation along PANTHER trees
- **Taxonomy:** taxonomy vs. phylogeny; Linnaean ranks vs. clades and why ranks are not comparable across groups; codes of nomenclature, *Candidatus* names and the SeqCode; species concepts, DNA–DNA hybridisation, 16S thresholds and ANI ≈ 95%; GTDB, monophyly and relative evolutionary divergence; NCBI Taxonomy as a working classification; sequence-based taxonomy of uncultivated organisms
- **Tree of life:** Woese and the three domains; candidate phyla radiation; GTDB; rooting with ancient paralogues; LUCA; Asgard archaea and the two-domain tree; eukaryotic supergroups; tree vs. network of life
- **Hominoid evolution:** the great-ape tree; divergence times and their dependence on mutation rate; incomplete lineage sorting in the human–chimpanzee–gorilla trio; CoalHMM; the hominin fossil record

4.7 Recommended reading

These papers are recommended for reading in full:

- Saitou and Nei (1987) — neighbour joining; the distance method still used everywhere, and a model of a clear algorithm paper.
- Felsenstein (1981) — maximum likelihood trees from DNA sequences and the pruning algorithm.
- Kapli et al. (2020) — tree building in the genomic age; an up-to-date review of methods, models and pitfalls.
- Minh, Schmidt, et al. (2020) — IQ-TREE 2; what a modern maximum-likelihood program does (models, search, support).
- Philippe et al. (2011) — why more sequences are not enough; systematic error in phylogenomics.
- Hug et al. (2016) — a new view of the tree of life from thousands of genomes.

Titles, abstracts, links and the full texts are in Section A.4; a longer list of further papers is in Section B.4.

4.8 Background reading (open access)

- **Understanding Evolution: Phylogenetic systematics** — An illustrated introduction from Berkeley's Understanding Evolution site to phylogenetic trees: how to read them, how they are built with cladistics, and common misconceptions.
  https://evolution.berkeley.edu/phylogenetic-systematics/
- **Phylogenetics (Wikipedia)** — In biology, phylogenetics is the study of the evolutionary history of life using observable characteristics of organisms, which is known as phylogenetic inference.
  https://en.wikipedia.org/wiki/Phylogenetics
- **Phylogenetic tree (Wikipedia)** — A phylogenetic tree or phylogeny is a graphical representation which shows the evolutionary history between a set of species or taxa during a specific time.
  https://en.wikipedia.org/wiki/Phylogenetic_tree
- **Cladistics (Wikipedia)** — Cladistics is an approach to biological classification in which organisms are categorized in groups ("clades") based on hypotheses of most recent common ancestry.
  https://en.wikipedia.org/wiki/Cladistics
- **Computational phylogenetics (Wikipedia)** — Computational phylogenetics, phylogeny inference, or phylogenetic inference focuses on computational and optimization algorithms, heuristics, and approaches involved in phylogenetic analyses. The goal is to find a phylogenetic tree representing optimal evolutionary ancestry between a set of genes, species, or taxa.
  https://en.wikipedia.org/wiki/Computational_phylogenetics
- **Models of DNA evolution (Wikipedia)** — A number of different Markov models of DNA sequence evolution have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. These models are frequently used in molecular phylogenetic analyses.
  https://en.wikipedia.org/wiki/Models_of_DNA_evolution
- **Substitution model (Wikipedia)** — In biology, a substitution model, also called models of sequence evolution, are Markov models that describe changes over evolutionary time. These models describe evolutionary changes in macromolecules, such as DNA sequences or protein sequences, that can be represented as a sequence of symbols.
  https://en.wikipedia.org/wiki/Substitution_model
- **UPGMA (Wikipedia)** — UPGMA is a simple agglomerative (bottom-up) hierarchical clustering method. It also has a weighted variant, WPGMA, and they are generally attributed to Sokal and Michener.
  https://en.wikipedia.org/wiki/UPGMA
- **Neighbor joining (Wikipedia)** — In bioinformatics, neighbor joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees, created by Naruya Saitou and Masatoshi Nei in 1987.
  https://en.wikipedia.org/wiki/Neighbor_joining
- **Maximum parsimony (Wikipedia)** — In evolutionary biology, maximum parsimony is a method used to build phylogenetic trees based on the principle of simplicity. Under the maximum-parsimony criterion, the best tree will minimize the amount of homoplasy.
  https://en.wikipedia.org/wiki/Maximum_parsimony
- **Bayesian inference in phylogeny (Wikipedia)** — Bayesian inference of phylogeny combines the information in the prior and in the data likelihood to create the so-called posterior probability of trees, which is the probability that the tree is correct given the data, the prior and the likelihood model.
  https://en.wikipedia.org/wiki/Bayesian_inference_in_phylogeny
- **Bootstrapping (statistics) (Wikipedia)** — Bootstrapping is a procedure for estimating the distribution of an estimator by resampling one's data or a model which is estimated from the data. Bootstrapping assigns measures of accuracy to sample estimates.
  https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

- **Phylogenomics (Wikipedia)** — Phylogenomics is the intersection of the fields of evolution and genomics. The term has been used in multiple ways to refer to analysis that involves genome data and evolutionary reconstructions. It is a group of techniques within the larger fields of phylogenetics and genomics.
  https://en.wikipedia.org/wiki/Phylogenomics
- **Incomplete lineage sorting (Wikipedia)** — Incomplete lineage sorting (ILS) (also referred to as hemiplasy, deep coalescence, retention of ancestral polymorphism, or trans-species polymorphism) is a phenomenon in evolutionary biology and population genetics that results in discordance between species and gene trees.
  https://en.wikipedia.org/wiki/Incomplete_lineage_sorting
- **Phylogenetic network (Wikipedia)** — A phylogenetic network is any graph used to visualize evolutionary relationships between nucleotide sequences, genes, chromosomes, genomes, or species. They are employed when reticulation events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved.
  https://en.wikipedia.org/wiki/Phylogenetic_network
- **Tree of life (biology) (Wikipedia)** — The tree of life or universal tree of life is a metaphor, conceptual model, and research tool used to explore the evolution of life and describe the relationships between organisms, both living and extinct, as described in a famous passage in Charles Darwin's On the Origin of Species (1859).The affinities of all the beings of the same class have sometimes been represented by a great tree.
  https://en.wikipedia.org/wiki/Tree_of_life_(biology)
- **Two-domain system (Wikipedia)** — The two-domain system is a biological classification system that groups all cellular organisms in the tree of life into two domains: Archaea and Bacteria, while eukaryotes are considered a specialized clade under Archaea.
  https://en.wikipedia.org/wiki/Two-domain_system
- **Human evolution (Wikipedia)** — Homo sapiens is a distinct species of the hominid family of primates, which includes all the great apes. Over their evolutionary history, humans gradually developed traits such as bipedalism, dexterity, and complex language.
  https://en.wikipedia.org/wiki/Human_evolution

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

4.9 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

4.10 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Phylogenetics" and "Phylogenomics and Discordant Histories" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

4.10.1 Interpreting tree topology

In a published tree, human is drawn next to mouse, and chimpanzee further away. Does the drawing show that humans are more closely related to mice than to chimpanzees? Justify your interpretation.

Extensions, solution outline and common errors

Extensions:

1. What information on a tree defines relatedness?
2. What changes when branches are rotated?
3. What is the difference between a rooted and an unrooted tree?
4. Can a living species at a tip be the ancestor of another tip?

Solution outline:

- Ignore the distance on the page and the order of the tips; find the most recent common ancestors implied by the topology.
- Rotating the descendants around an internal node does not change the topology.
- On a rooted tree, compare how recent the shared ancestors are; on an unrooted tree, only splits can be discussed, without direction.
- State how the tree was rooted and whether that assumption is justified.
- Distinguish branch lengths, support values and time.
- Sampled living species are treated as tips, not as the ancestors of other tips.

Common errors:

- Reading the left-to-right order of the tips, or their closeness on the page, as relatedness.
- Calling the first-diverging living lineage “primitive”.
- Treating the arbitrary root of a plotting program as biological evidence.

Connections: Section 4.2; homology in 1; orthology in Section 3.4.

4.10.2 Comparing parsimony, distance, likelihood and Bayesian inference

An aligned single-copy gene has been sequenced from 40 species. Choose and justify a strategy for building the tree, including preprocessing, the estimation of uncertainty and checks of the model.

Extensions, solution outline and common errors

Extensions:

1. How does neighbour joining differ from maximum likelihood?
2. Why is parsimony not free of assumptions?
3. What extra quantity does Bayesian inference compute?
4. Is the likelihood of the best tree the probability that the tree is true?

Solution outline:

- The target is a gene tree. Check the alignment, recombination and orthology first.
- A neighbour-joining tree from corrected distances is a fast first look, but reducing the alignment to pairwise distances discards information in the site patterns.

- Parsimony minimises the number of character changes and can be statistically inconsistent when branch lengths are very unequal (long-branch attraction).
- Maximum likelihood maximises $P(D \mid T, b, \theta)$ under an explicit substitution model, computed with the pruning algorithm, and searches tree space heuristically.
- Bayesian inference combines the likelihood with priors to sample from $P(T, b, \theta \mid D)$, using MCMC, and requires checks of convergence.
- For a single locus, model-based maximum likelihood with several searches, support values and model checks is a reasonable default; another approach is needed if the history violates the assumption of a single tree.

Common errors:

- Saying that maximum likelihood "calculates all possible trees".
- Confusing $P(D \mid T)$ with $P(T \mid D)$.
- Choosing the software before checking what history the sequences represent.

Connections: Section 4.3; Section 4.3.9; multiple alignment in Section 1.6.

4.10.3 Substitution models and likelihood

Explain why the proportion of differing sites is generally not a good measure of evolutionary distance, and describe the role of the substitution model in the likelihood of a tree.

Extensions, solution outline and common errors

Extensions:

1. What is a rate matrix?
2. Compare JC69, HKY85 and GTR.
3. What does gamma-distributed rate variation do?
4. How can a likelihood be computed without assigning every ancestral sequence?

Solution outline:

- Observed differences hide multiple substitutions and reversions at the same site; evolutionary distance is the expected number of substitutions per site.
- Define the rate matrix Q and $P(t) = e^{Qt}$; the branch length connects the process to the transition probabilities.
- The models differ in their assumptions about equilibrium frequencies and exchange rates; a more complex model is not automatically closer to the truth.
- Discrete gamma categories approximate variation in rate between sites; they do not model changes in the rate of a site over time.
- Felsenstein's pruning algorithm propagates conditional likelihood vectors from the leaves towards the root, multiplying over the child branches at each internal node and summing over the possible states at the root.
- Pruning makes the likelihood of one tree fast to compute; the search over topologies remains a hard combinatorial problem.

Common errors:

- Calling substitutions directly observed events.
- Saying that GTR is always best because it has more parameters.
- Treating the arbitrary root used in the pruning calculation as a biological root.

Connections: Section 4.3.6; Section 4.3; mutation versus substitution in Section 3.3; alignment scores in Section 1.3.

4.10.4 A strongly supported but implausible tree

Two rapidly evolving parasites form a clade with 100% bootstrap support, contrary to well-supported independent evidence. Design an analysis that distinguishes random from systematic error.

Extensions, solution outline and common errors

Extensions:

1. What is long-branch attraction?
2. Can maximum likelihood also be wrong?
3. Would removing the fast-evolving taxa solve the problem?
4. What does 100% bootstrap support actually mean?

Solution outline:

- Recheck the identity of the sequences, contamination, orthology, domain architecture, taxon coverage and alignment.
- In long-branch attraction, convergent changes on two separate long branches outweigh the signal of a short internal branch; parsimony can then be statistically inconsistent.
- Maximum likelihood can also fail when the model is badly misspecified, for example with heterogeneous composition or rates, or when the taxon sampling is poor.
- Test for compositional bias and saturation, add taxa that break up the long branches, vary the outgroups, and use site- or composition-heterogeneous models.
- Compare support across alignments and models rather than deleting inconvenient taxa.
- A bootstrap value of 100% means that the split is recovered from every resampled alignment with the same method; it is not the probability that the split is true, and it does not protect against errors shared by all replicates.

Common errors:

- Diagnosing long-branch attraction merely because two branches look long.
- Assuming that more sites remove systematic error.
- Removing discordant taxa until the expected tree is recovered.

Connections: Section 4.3; Section 4.3.9; systematic error in Section 4.4.

4.10.5 Rooting and dating a tree

Given an unrooted gene tree of mammals, explain how to place the root and estimate divergence times with their uncertainty.

Extensions, solution outline and common errors

Extensions:

1. What makes a good outgroup?
2. Why can midpoint rooting fail?

3. What information separates rate from time?
4. Compare strict and relaxed clocks.

Solution outline:

- Check that the gene history is appropriate for the question about species, and that the outgroup sequences are homologous.
- Use independently justified outgroups that are not too distant, and test whether the root is stable across outgroups and models.
- Midpoint rooting assumes roughly equal distances from the root to all tips; variation in rate between lineages violates this.
- A branch length is $b = r\Delta t$: sequences estimate the amount of change, while dated tips, fossils, geological events or an external rate give the time scale.
- A strict clock assumes one rate for all branches; relaxed clocks allow the rate to vary between branches.
- Specify where the calibrations are placed and their uncertainty, the tree prior, and sensitivity analyses; report credible intervals that are conditional on these choices.

Common errors:

- Treating the most "primitive-looking" or longest-branched taxon as the root.
- Equating the age of a fossil with the exact age of a node.
- Reporting precise dates without calibrations or sensitivity to the priors.

Connections: Section 4.2; Bayesian dating in Section 4.3; the molecular clock in Section 3.3.

4.10.6 A species tree under incomplete lineage sorting

Hundreds of gene trees disagree about the relationships of three species from a rapid radiation. Does this refute a bifurcating species history? Design an appropriate analysis.

Extensions, solution outline and common errors

Extensions:

1. What is incomplete lineage sorting?
2. Give the probabilities of the three gene trees for three species.
3. Why can a majority vote over gene trees fail with more species?
4. Compare concatenation, ASTRAL and full multispecies-coalescent methods.

Solution outline:

- In incomplete lineage sorting, ancestral gene lineages persist through two successive speciations before they coalesce.
- With the internal branch of the species tree of length t in coalescent units, the gene tree that matches the species tree has probability $1 - \frac{2}{3}e^{-t}$, and each of the two discordant trees $\frac{1}{3}e^{-t}$.
- Short internal branches and large ancestral effective population sizes increase discordance.
- With more species, a discordant gene tree can be more frequent than the matching one (the anomaly zone), so a simple majority vote can give the wrong species tree (Degnan and Rosenberg 2006).
- Concatenation assumes one tree for all genes. Summary methods such as ASTRAL use the

estimated gene trees, and full methods also integrate over the uncertainty in sequences and gene trees.
- Check for gene-tree estimation error, recombination within loci, hidden paralogy and introgression before attributing all conflict to incomplete lineage sorting.

Common errors:

- Saying that every locus should match the species tree.
- Treating loci as independent without checking linkage and recombination.
- Interpreting an ASTRAL tree as robust to every kind of discordance.

Connections: Section 4.4; drift and effective population size in Section 3.3; great-ape evolution in Section 4.5.

4.10.7 Recombination, horizontal transfer and other causes of conflict

A bacterial gene supports a topology different from the tree of the core genome. Evaluate horizontal transfer as one hypothesis among the biological and analytical explanations of the conflict.

Extensions, solution outline and common errors

Extensions:

1. How does recombination violate standard tree inference?
2. How could recombination breakpoints be detected?
3. What else causes conflict between gene trees and species trees?
4. When is a network preferable to a tree?

Solution outline:

- Horizontal transfer is a hypothesis, not the definition of conflict.
- Check orthology, contamination, assembly context, alignment, the amount of signal, composition and support values.
- Recombination changes the tree along the sequence; detect breakpoints and analyse the segments separately, or use a method that allows recombination, rather than forcing one tree.
- Look for independent evidence of transfer: a well-supported conflicting gene tree, an insertion at an unusual genomic position, a patchy presence–absence pattern, and composition compared with suitable controls.
- Consider incomplete lineage sorting, duplication and loss, introgression, recombination and estimation error as alternatives.
- Use reconciliation or explicit network models when reticulation is the target; a split network that only displays incompatible signals is descriptive.

Common errors:

- Calling every discordant bacterial gene horizontally transferred.
- Treating unusual GC content as proof.
- Building one tree from a mosaic alignment and taking its support values literally.

Connections: Section 4.4; horizontal gene transfer in Section 3.7; gene order in Section 6.5; pan-genomes in Section 6.6.

4.10.8 Designing a phylogenomic analysis

Design a study of the history of a difficult radiation of 80 species with 2,000 candidate loci, including alternative representations if a single bifurcating species tree is inadequate.

Extensions, solution outline and common errors

Extensions:

1. How would you choose taxa and loci?
2. Would you trim the alignments and partition the data by gene?
3. Which results would show that the conclusions are robust?
4. What must be archived so that the analysis can be repeated?

Solution outline:

- Define whether the target is a bifurcating species tree, a reticulate history or the distribution of gene histories.
- Sample taxa that break up long branches; establish orthology with gene trees, copy number, conserved gene order and domains; record missing data.
- Align each locus, inspect the uncertainty in the alignments, define loci so that they do not span recombination breakpoints, and trim moderately rather than automatically.
- Compare partitioning and model schemes; infer the gene trees with their uncertainty, and a partitioned concatenated maximum-likelihood tree.
- Estimate a coalescent species tree, report gene and site concordance factors, and test for introgression or duplication and transfer where the conflict is structured.
- Test the robustness of the topology to alignments, models, taxon subsets, loci, outgroups and support measures.
- Archive the accession numbers and versions, alignments, exclusions, partitions, commands, software versions, random seeds, trees and bootstrap or posterior samples.

Common errors:

- Assuming that 2,000 loci guarantee the correct answer.
- Reporting only a consensus tree with support values and hiding the discordance.
- Letting the preferred topology decide which data are filtered out.

Connections: Section 4.4; orthology inference in Section 3.4; alignment uncertainty and support in Section 4.3.9.

4.10.9 Data type, saturation and the clock

You want to date the divergence of several vertebrate classes from a set of nuclear protein-coding genes. Decide whether to analyse nucleotides, codons or amino acids, and how you would test whether a molecular clock can be assumed.

Extensions, solution outline and common errors

Extensions:

- How would you detect saturation of third codon positions?
- How can differences in GC content between lineages mislead a tree?
- Give the degrees of freedom of the likelihood-ratio test of a clock for n sequences, and explain them.

Solution outline:

1. At divergences of hundreds of millions of years, third codon positions are usually saturated: plot uncorrected against corrected distances, or compare trees from different codon positions.
2. Use amino acids, or nucleotides without third positions (or RY-coded), possibly partitioned by codon position; check for compositional heterogeneity and consider non-stationary or recoded analyses; codon models are realistic but costly.
3. Test the clock with relative rate tests for pairs of lineages against an outgroup, and with a likelihood-ratio test of the clock-constrained tree ($n - 1$ node heights) against the unconstrained tree ($2n - 3$ branch lengths), $n - 2$ degrees of freedom.
4. If the strict clock is rejected, as is likely, use a relaxed-clock Bayesian analysis with several fossil calibrations, and report how sensitive the dates are to the clock model and the calibrations.

Common errors:

- Using all nucleotide positions for deep divergences and trusting high support.
- Treating branch lengths as times without a clock model.
- Assuming that failure to reject the clock proves it (low power with few sites).

Connections: Section 4.3.6, Section 4.3; the molecular clock (Section 3.3); detecting selection with codon models (Section 3.6).

5 Genomics

> **From the exam description**
>
> Methods for sequencing, algorithms for assembly, gene predictions, etc.

5.1 Overview

Genomics turns DNA molecules into genome sequences and genome sequences into biological knowledge. In less than fifty years, sequencing has gone from a few hundred bases per experiment to complete human genomes in a day, and each new technology has required new algorithms. This chapter follows that history, with the focus on the computational methods. Students should know how the technologies work, but above all why they lead to the algorithmic problems they do.

- sequencing technologies, from Sanger to single-molecule long reads, and the theory of shotgun sequencing and coverage (Section 5.2);
- genome assembly: overlap graphs, string graphs and de Bruijn graphs, repeats, scaffolding and assembly quality (Section 5.3);
- read mapping and variant calling: indexes, SAM and VCF records, mapping quality, genotype likelihoods, structural variants, phasing, targeted sequencing, copy number and somatic variants in cancer (Section 5.4);
- regulatory and epigenomic sequencing: ChIP-seq, ATAC-seq, bisulfite and long-read methylation calling, and CLIP-seq (Section 5.5);
- gene prediction in prokaryotes and eukaryotes (Section 5.6);
- metagenomics: amplicon and shotgun sequencing, profiling, assembly and binning (Section 5.7);
- human genetics: linkage disequilibrium, genome-wide association studies, fine-mapping, polygenic scores, and the search for Mendelian disease variants by exome and genome sequencing (Section 5.8).

The algorithms for exact matching that underlie read mapping are described in Section 1.4. What genome sequences have revealed about genome organisation and evolution is the subject of Chapter 6, and the sequencing of RNA of Chapter 7.

5.2 Sequencing methods

Every genome analysis starts with reading DNA, and the properties of the reads, their length, their error profile and their cost, determine which algorithms are needed. This section describes the main sequencing technologies, roughly in historical order, and the basic theory of shotgun sequencing (Shendure et al. 2017; Goodwin et al. 2016).

5.2.1 First generation: Sanger sequencing

Two methods for sequencing DNA were published in 1977. Allan Maxam and Walter Gilbert used chemicals that cleave DNA after specific bases (Maxam and Gilbert 1977). Frederick Sanger's **chain-termination method** (Sanger, Nicklen, et al. 1977) became the standard. A DNA polymerase copies a single-stranded template in the presence of normal nucleotides and a small amount of *dideoxynucleotides* (ddNTPs), which lack the 3' hydroxyl group needed to extend the chain. Each time a ddNTP is incorporated, synthesis stops, so the reaction produces fragments ending at every position. Originally four reactions, one per base, were run in separate lanes of a gel. The same year Sanger's group published the first complete DNA genome, the 5,386 bases of the bacteriophage φX174, and in 1980 Sanger and Gilbert shared the Nobel Prize in Chemistry.

Sanger sequencing was automated in the 1980s by labelling each of the four ddNTPs with a different fluorescent dye, so that all four reactions could run in one lane and be read by a laser (Smith et al. 1986), and later by replacing gels with capillaries. Capillary sequencers produced the first bacterial genomes and the human genome. Reads are 500–1,000 bases long and very accurate. To use them, each base needed a reliable error estimate. The program **Phred** (Ewing et al. 1998) introduced the quality score; most modern sequencing technologies still report Phred-scaled or Phred-like base qualities,

$$Q = -10 \log_{10} P_{\text{error}},$$

so that $Q = 20$ means a 1% error probability and $Q = 30$ 0.1%. Sanger sequencing is still used to check individual clones and mutations, but it is too slow and expensive for genomes: it produces about one read per capillary per run.

5.2.2 Second generation: massively parallel short reads

"Next-generation" sequencing (NGS) replaced the separation of fragments by size with *sequencing by synthesis* of millions of immobilised, clonally amplified DNA molecules, imaged in parallel.

- **Pyrosequencing** detects the pyrophosphate released when a nucleotide is incorporated, through an enzymatic cascade that ends with light from luciferase (Ronaghi et al. 1998). One nucleotide type is added at a time. **454 sequencing** (2005) ran pyrosequencing in about a million picolitre wells, each containing a bead with DNA amplified by emulsion PCR (Margulies et al. 2005). Reads were 100–700 bases. Since the signal is proportional to the number of identical bases incorporated, the characteristic errors are **indels in homopolymer runs**.
- **Illumina** sequencing (originally Solexa) amplifies each molecule into a cluster on a glass flow cell by *bridge amplification*, and then adds all four nucleotides at once, each with a different fluorescent label and a reversible terminator, so that exactly one base is added per cycle (Bentley et al. 2008). After imaging, the label and terminator are removed and the next cycle begins. Reads are typically 2×150 bases. Errors are mainly **substitutions**, at a rate typically well below 1% (depending on instrument and chemistry), increasing towards the end of the read as molecules in a cluster get out of phase. Illumina came to dominate the market. It reduced the cost of a human genome from millions of dollars in 2007 to around a thousand dollars in 2014 and a few hundred in 2026.
- **Ion Torrent** sequencing detects the protons released during incorporation with a semiconductor chip, a pH meter rather than a camera. Like 454, it has homopolymer errors.

Short reads are cheap and accurate, but they cannot span repeats longer than a few hundred bases, which limits both assembly (Section 5.3) and the detection of structural variants (Section 5.4).

5.2.3 Third generation: single-molecule long reads

Single-molecule technologies read native DNA molecules without amplification, and produce reads that are tens of kilobases or longer.

- **Pacific Biosciences** (PacBio) watches a single DNA polymerase at the bottom of a *zero-mode waveguide*, a well so small that only the fluorescent nucleotide being incorporated is illuminated (Eid et al. 2009). The original "continuous long reads" (CLR) had error rates of 10–15%, mostly random insertions and deletions. In **HiFi** sequencing, a circular template is read many times, and a consensus of the passes (*circular consensus sequencing*) gives reads of 15–25 kb with an accuracy above 99.9% (Wenger et al. 2019).
- **Oxford Nanopore** (ONT) passes a single DNA strand through a protein pore in a membrane, and measures the changes in the ionic current, which depend on the bases currently in the pore (Deamer et al. 2016). The signal is translated into bases by a neural-network *basecaller*. Read length is limited only by the length of the DNA molecule. "Ultra-long" reads can exceed a megabase (M. Jain et al. 2018). Error rates were initially 10–15%, with systematic errors in homopolymers. With newer pores, chemistry and basecallers they have fallen to around 1% for single-strand (*simplex*) reads in 2026, and lower for *duplex* reads, in which both strands of a molecule are read. Nanopore devices are small, sometimes handheld, and give results in real time, which has made them useful in the field and in outbreak surveillance. ONT can also sequence RNA directly.

Quoted accuracies change quickly and depend on the chemistry, the basecaller, the read type (CLR or HiFi, simplex or duplex) and whether a consensus of several reads is taken. Compare technologies qualitatively, by their error *types*, and treat any number as a snapshot with a date.

Both long-read technologies also detect **base modifications**, such as cytosine methylation, from the polymerase kinetics (PacBio) or the current signal (ONT), without chemical conversion.

Long-read transcriptomics. Short RNA-seq reads cover only fragments of a transcript, so which exons belong together in one isoform must be inferred, which is ambiguous for genes with many isoforms. Long reads can cover a complete transcript in a single read. PacBio **Iso-Seq** sequences full-length cDNA as HiFi reads. ONT sequences either cDNA or, in **direct RNA sequencing**, the native RNA molecule itself, without reverse transcription or PCR (Garalde et al. 2018), so that RNA modifications such as N^6-methyladenosine, and the length of the poly(A) tail, can in principle be inferred directly from the electrical signal. Modification calling remains dependent on the model, the sequence context and the training data. Direct RNA sequencing needs more input RNA and has higher error rates than DNA sequencing. Since long-read runs give fewer reads than short-read runs, they are mainly used to discover and annotate isoforms (Section 5.6), while expression levels are still often quantified with short reads. The analysis of RNA-seq data is described in Section 7.3.

Table 5.1: Main sequencing technologies. Numbers are approximate, as of 2026, and change quickly.

Technology	Read length	Main error type	Typical accuracy	Strengths
Sanger (capillary)	500–1,000 bp	few	> 99.9%	accurate, simple; low throughput
454 pyrosequencing	100–700 bp	homopolymer indels	~99%	first NGS; discontinued
Illumina	2 × 50–300 bp	substitutions	> 99.9%	cheap, very high throughput
PacBio HiFi	15–25 kb	few, some indels	> 99.9%	long and accurate
Oxford Nanopore	10 kb – > 1 Mb	indels, homopolymers	~99% (simplex), higher for duplex	longest reads, portable, real time

5.2.4 Base qualities are model estimates

A base call and its quality are the output of a statistical model, the basecaller, applied to an optical or electrical signal. A Phred score is therefore a *predicted* error probability. It is useful only if it is *calibrated*: among all bases given $Q = 30$, about one in a thousand should be wrong when compared with a trusted sequence. Calibration is checked empirically, and base-quality recalibration (Section 5.4) corrects it using known variant sites. Qualities from different technologies, chemistries and basecaller versions are not directly comparable, and re-basecalling the same raw signal with a newer model can change the reads. The basecaller and its model version are therefore part of the record of an experiment.

Errors are also not independent. Consensus methods, such as PacBio HiFi reads built from several passes over the same molecule, or the consensus of many reads in an assembly, remove random errors efficiently, but *systematic* errors that depend on the sequence context, such as indels in homopolymers, recur in every read and are not averaged away. Residual errors in high-accuracy data are therefore concentrated in particular sequence contexts.

5.2.5 Shotgun sequencing and coverage

Because each read covers only a small part of a genome, genomes are sequenced by the **shotgun** method: the DNA is broken into random fragments, many of them are sequenced, and the genome is reconstructed computationally from the overlaps. Rodger Staden described the computational strategy in 1979 (Staden 1979). Early genome projects first mapped large clones and then sequenced each clone by shotgun (*clone-by-clone*), while *whole-genome shotgun* sequencing was first used for the bacterium *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995). The debate about whether whole-genome shotgun could work for the human genome (Weber and Myers 1997) is described in Chapter 6.

How much sequencing is needed? If N reads of length L are sampled uniformly from a genome of length G, the **coverage** (depth) is

$$c = \frac{NL}{G}.$$

Lander and Waterman showed in 1988 that the number of reads covering a given base is then approximately Poisson-distributed with mean c, so the fraction of the genome not covered by any read is e^{-c} (Lander and Waterman 1988). With 8× coverage this is about 0.03%, or about 1 kb in a 3 Mb bacterial genome. The expected number of contigs, islands of overlapping reads, is about $Ne^{-c(1-\theta)}$, where θ is the fraction of a read that must overlap with another to be detected. In practice, coverage is far from uniform because of biases in library preparation and amplification, especially in GC-rich or GC-poor regions, so much higher depths are used: 30× is standard for human resequencing with short reads, and 50–100× for assembly.

5.2.6 Paired-end and mate-pair reads

Read pairs add long-range information to short reads. In **paired-end** sequencing, both ends of each fragment of a few hundred bases are read, so the two reads are known to lie at an approximately known distance, in opposite orientation. Two lengths must be kept apart. The **read length** is the number of bases sequenced from each end (for example 150). The **insert size** (or fragment length) is the length of the DNA fragment between the adapters, from the outer end of one read to the outer end of the other (for example 300–500 bases); it is not chosen by the sequencer but by library preparation, and it varies around a mean. If the insert is shorter than twice the read length the two reads overlap, and if it is shorter than one read the read continues into the adapter, which must then be trimmed. **Mate-pair** libraries

are made by circularising longer fragments (2–20 kb) and sequencing across the junction, giving pairs that span much larger distances. Pairs are used in assembly to order and orient contigs (*scaffolding*), in mapping to place reads in repeats, and in variant calling to detect structural variants, whose signature is pairs that map at unexpected distances or orientations. Other sources of long-range information are *Hi-C* chromatin contact data, optical maps of long DNA molecules, and, increasingly, long reads themselves.

See also: further explanation in Section D.5.1; related Wikipedia articles in Section E.5.1.

Sources

Parts of this section are adapted from the Wikipedia articles DNA sequencing, Sanger sequencing, Massive parallel sequencing, Illumina dye sequencing, Single-molecule real-time sequencing, Nanopore sequencing, Phred quality score, Shotgun sequencing and Coverage (genetics) (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Sequencing, Mapping, Assembly, and Variation" (CC BY 4.0), used with permission.

5.3 Genome assembly

De novo assembly reconstructs a genome from reads without a reference. It is one of the classical algorithmic problems of bioinformatics, and the algorithms have changed with each generation of sequencing technology (Nagarajan and Pop 2013). This section describes the main paradigms, the central role of repeats, scaffolding and the evaluation of assemblies.

5.3.1 The problem and why repeats matter

If reads were error-free and the genome had no repeats, assembly would be the **shortest common superstring** problem: find the shortest string that contains every read. This problem is NP-hard, but a greedy algorithm that repeatedly merges the two reads with the longest overlap works well in practice, and the first assemblers, such as phrap and the TIGR Assembler used for *H. influenzae* (Fleischmann et al. 1995), were essentially greedy.

The real difficulty is **repeats**. If a repeat is longer than the reads (or than the span of read pairs), the reads from its copies cannot be told apart. The assembler cannot know how the unique sequences on either side of the copies are connected, and the shortest superstring would collapse the copies into one. Repeats are everywhere: transposable elements make up about half of the human genome, and segmental duplications, tandem repeats, ribosomal RNA operons and centromeric satellites span tens of kilobases to megabases. An assembler therefore stops at repeats it cannot resolve and outputs **contigs**, contiguous sequences between unresolved repeats or gaps in coverage. Modern assemblers represent the ambiguity explicitly as a graph, and the history of assembly is largely a history of how to build and simplify such graphs (Compeau et al. 2011).

5.3.2 Overlap–layout–consensus and string graphs

The **overlap–layout–consensus** (OLC) paradigm, formalised for shotgun sequencing in the 1980s and 1990s, has three steps (Anson and Myers 1999):

1. **Overlap**: find all pairs of reads that overlap, usually by first finding shared k-mers (seeds) and then aligning the candidate pairs. With N reads this is the most expensive step.

2. **Layout**: build an *overlap graph* with reads as nodes and overlaps as edges, and find paths through it that correspond to contigs. Ideally one wants a path that visits every read exactly once, a *Hamiltonian path*, which is NP-hard in general.
3. **Consensus**: align the reads in each contig and compute the consensus sequence, which corrects sequencing errors.

OLC assemblers built the first large genomes from Sanger reads: the Celera Assembler for *Drosophila* (Myers et al. 2000) and the human genome, Arachne and others. Myers simplified the overlap graph to the **string graph** (Myers 2005) by removing contained reads and *transitive* edges (if A overlaps B, B overlaps C, and A overlaps C, the edge A–C is redundant). In the string graph, unambiguous stretches collapse into chains (*unitigs*), and branching points correspond to repeat boundaries.

5.3.3 De Bruijn graphs

With second-generation sequencing, assembly had to handle hundreds of millions of short reads, and computing all pairwise overlaps became too expensive. The solution was a graph formulation originally proposed for sequencing by hybridisation (Idury and Waterman 1995; Pevzner et al. 2001). Each read is broken into all its overlapping ***k*-mers**, and a **de Bruijn graph** is built with the $(k-1)$-mers as nodes and each k-mer as an edge from its prefix to its suffix. For the sequence `ATGGCGTGCA` and $k = 3$, for example, the edges are `ATG`, `TGG`, `GGC`, `GCG`, `CGT`, `GTG`, `TGC`, `GCA`, connecting the 2-mers `AT` $\to$ `TG` $\to$ `GG` $\to$ `GC` $\to$ `CG` $\to$ `GT` $\to$ `TG` $\to$ `GC` $\to$ `CA`. The repeated 2-mers `TG` and `GC` become single nodes visited twice, so the graph contains a cycle.

The genome is a path that uses every *edge* once, an **Eulerian path**, which can be found in linear time. This is the key algorithmic advantage over the Hamiltonian path of the overlap graph (Compeau et al. 2011). In practice, an exact Eulerian path is not what assemblers compute. Repeats longer than k create branches in the graph, and the assembler outputs the unambiguous paths between branches as contigs. Sequencing errors create characteristic structures that must be removed: *tips* (dead ends caused by an error near the end of a read) and *bubbles* (two parallel paths caused by an error in the middle of a read, or by a heterozygous site). Low-coverage k-mers are discarded or corrected before the graph is built.

The choice of k is a trade-off: a small k makes the graph more tangled by repeats, a large k makes it fragmented where coverage is low. Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009) and SOAPdenovo were among the first short-read de Bruijn assemblers. SPAdes (Bankevich et al. 2012) combines graphs for several values of k and handles the uneven coverage of single-cell and metagenomic data, and MEGAHIT (Li et al. 2015) uses succinct data structures to reduce memory.

Short-read assemblies of large eukaryotic genomes are fragmented into tens of thousands of contigs, because most repeats are longer than the reads. Such draft genomes contain most genes, but repeats, segmental duplications and centromeres are collapsed or missing.

5.3.4 Long-read assembly

Long reads span most repeats and have brought assembly back to overlap-based methods. The first long-read assemblies corrected the noisy PacBio reads with each other before assembly (Chin et al. 2013). Canu adapted OLC to noisy reads (Koren et al. 2017), and Flye builds a *repeat graph* from approximate overlaps (Kolmogorov et al. 2019). With accurate HiFi reads, string-graph assemblers such as **hifiasm** (Cheng et al. 2021) produce contigs that are often tens of megabases long. Today, the best assemblies combine HiFi reads for accuracy, ultra-long nanopore reads to span the longest repeats, and Hi-C or parental data to separate haplotypes and chromosomes.

The contrast "de Bruijn graphs for short reads, overlap graphs for long reads" is a useful way to understand the two ideas, and describes the history well, but it is not a classification of current assemblers. Many modern tools mix the ideas. Flye's repeat graph is a de Bruijn-like graph built from approximate overlaps, several HiFi assemblers build de Bruijn or related graphs from long k-mers or minimizers, and hybrid assemblers build a graph from accurate HiFi reads and resolve its tangles with ultra-long nanopore reads (Rautiainen et al. 2023). Most long-read tools find overlaps with minimizers (Section 5.4) rather than with all-against-all alignment.

Haplotype-resolved assembly. A diploid genome contains two copies of each chromosome, which differ at millions of sites. Traditional assemblers produced a *collapsed* consensus, a mosaic of the two haplotypes. With accurate long reads, the two haplotypes can be assembled separately (*phased*), either using reads from the parents to assign each read to a haplotype (*trio binning*) or using Hi-C data, which link variants on the same chromosome (Cheng et al. 2021).

This adds to the classical vocabulary of contigs and scaffolds. A *haplotig* is a contig that represents a single haplotype. A *collapsed* or *pseudo-haplotype* assembly may switch between haplotypes along a contig. Many assemblies are reported as a *primary* assembly, which contains one copy of each region, plus *alternate* haplotigs for the regions where the second haplotype differs. A *haplotype-resolved* (fully phased) assembly instead gives two complete sets of contigs, one per haplotype. A *chromosome-scale* assembly has scaffolds that each correspond to a chromosome but may still contain gaps, whereas a *telomere-to-telomere* assembly is gapless from one end of each chromosome to the other.

Telomere-to-telomere assemblies. In 2022 the Telomere-to-Telomere (T2T) consortium published the first complete sequence of a human genome, CHM13, from a cell line that is effectively haploid (Nurk et al. 2022). It added about 200 Mb (8%) to the previous reference, including all centromeres, the short arms of the acrocentric chromosomes and large segmental duplications. Complete genomes of other great apes followed (Yoo et al. 2025). The Vertebrate Genomes Project aims at reference-quality, near-complete assemblies of all vertebrate species (Rhie et al. 2021), and the Human Pangenome Reference Consortium has assembled phased genomes of many individuals to represent human diversity, rather than a single reference (Liao et al. 2023).

5.3.5 Polishing

The layout of an assembly fixes the order of the reads, but its consensus may still contain errors, particularly when it is built from noisy long reads. **Polishing** realigns reads to the draft contigs and recomputes the consensus. Racon, for example, builds a partial-order alignment of the reads mapped to each contig window (Vaser et al. 2017). Other polishers use the raw signal or models trained for one technology, and accurate short reads can correct remaining small errors in a long-read assembly. Polishing is usually repeated a few times, and should stop when independent measures of accuracy no longer improve.

Polishing can also make an assembly worse. Reads from one copy of a repeat may be mapped to another copy and make the copies identical; in a collapsed diploid assembly, true heterozygous sites are pushed towards the majority allele; and short-read polishing can reintroduce errors in regions that short reads map poorly. Polishing improves the local consensus but does not repair misjoins. For phased assemblies, reads should be assigned to a haplotype before polishing. Accuracy should be evaluated with data that were not used for polishing, for example with k-mers from an independent accurate read set, since evaluating with the same reads is circular.

5.3.6 Scaffolding

Scaffolding orders and orients contigs and estimates the size of the gaps between them, using information that spans gaps: read pairs, mate pairs, long reads, optical maps or Hi-C contacts, whose frequency decreases with distance along the chromosome, so that Hi-C can assign contigs to chromosomes and order them (Burton et al. 2013). The result is a set of **scaffolds** in which gaps are filled with stretches of `N`. Scaffolding is a graph problem: contigs are nodes, linking evidence gives edges, and conflicting links caused by repeats and chimeric reads must be resolved. Errors in scaffolding, such as inversions and misjoins, are a common source of assembly errors.

5.3.7 Assembly quality

Assemblies are evaluated by three properties:

- **Contiguity.** The **N50** is the length such that contigs of that length or longer contain half of the assembly. It is the most quoted statistic, but it is easy to inflate by joining contigs incorrectly. NG50 uses half of the estimated genome size instead of the assembly size, which makes assemblies of different sizes comparable. Contig and scaffold N50 should be reported separately.
- **Completeness. BUSCO** searches the assembly for a set of genes that are expected to be present in a single copy in a given lineage, and reports the percentage found complete, fragmented, duplicated or missing (Simão et al. 2015). *k*-mer based methods such as Merqury compare the *k*-mers in the assembly with those in accurate reads from the same individual (Rhie et al. 2020).
- **Correctness.** Base-level accuracy is measured as a consensus quality value (QV; Q40 means one error per 10 kb). Structural correctness (misjoins, inversions, collapsed repeats) is assessed by comparison with a reference, optical maps or Hi-C data.

Reference-based tools such as QUAST align an assembly to a related reference genome and report misassemblies, but differences from the reference can be real biological differences rather than errors (Gurevich et al. 2013). With *k*-mers from the parents, Merqury also measures how well the haplotypes of a phased assembly are separated (Rhie et al. 2020). Further aspects of quality, such as contamination and the representation of haplotypes, are summarised in Figure 5.1.

These are different axes. N50 measures only contiguity, not correctness: an assembler that joins contigs aggressively across repeats can produce a high N50 together with many misassemblies, and a complete set of BUSCO genes says nothing about the repeats between them. An assembly should therefore be judged on all three.

Community competitions compared assemblers on the same data. The Assemblathons showed that different assemblers excel by different metrics, and that the results vary strongly between genomes (Bradnam et al. 2013).

See also: further explanation in Section D.5.2; related Wikipedia articles in Section E.5.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Sequence assembly, De novo sequence assemblers, De Bruijn graph, Eulerian path, N50, L50, and related statistics, Scaffolding (bioinformatics) and Whole genome sequencing (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Sequencing, Mapping, Assembly, and Variation” (CC BY 4.0), used with permission.

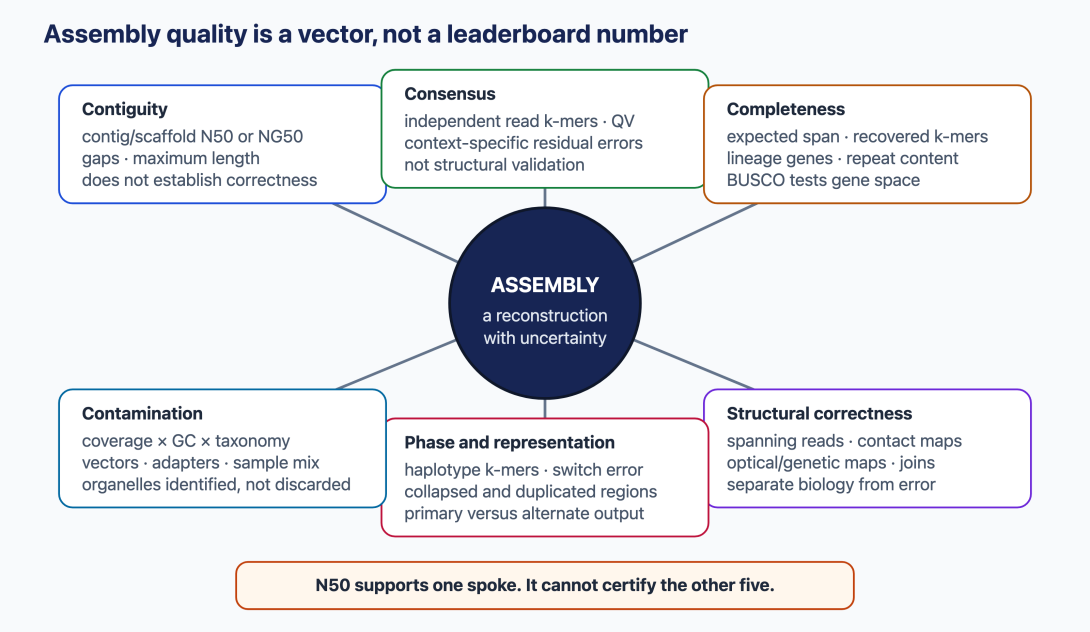


Figure 5.1: Assembly quality has several dimensions, each of which needs its own measure. N50 measures only contiguity. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

5.4 Mapping and variant calling

Once a reference genome exists, most sequencing is **resequencing**: reads from a new individual, tumour or strain are mapped to the reference, and the differences are identified. This is the basis of human genetics, cancer genomics, population genomics and clinical diagnostics.

5.4.1 Read mapping

A read mapper must place hundreds of millions of reads on a genome of billions of bases, allowing for sequencing errors and true differences, and decide how confident each placement is. Aligning each read with the Smith–Waterman algorithm against the whole genome is far too slow, so all mappers first **index** the reference, find candidate locations with exact matches (*seeds*), and then align the read only at those locations (*seed and extend*). The index structures are described in Section 1.4.

- **Hash-based mappers.** The first generation, such as MAQ (Li et al. 2008), built a hash table of short words in the reads or in the genome, often with *spaced seeds* to tolerate mismatches. They were fast but used much memory.
- **Burrows–Wheeler mappers. Bowtie** (Langmead et al. 2009) and **BWA** (Li and Durbin 2009) index the genome with the Burrows–Wheeler transform and the FM-index, which allows exact matches to be found in time proportional to the length of the read, independent of the genome size, with an index of a few gigabytes for the human genome. Mismatches are handled by backtracking. BWA-MEM, the standard for short reads, finds maximal exact matches as seeds and extends them with banded Smith–Waterman alignment.
- **Long-read mappers.** For long, noisy reads, **minimap2** (Li 2018) samples *minimizers* (the smallest k-mer in each window) as seeds, *chains* collinear seed hits by dynamic programming, and aligns only between the anchors. It has become the standard tool for long reads and genome-to-genome alignment.

RNA-seq reads must be aligned across introns with *spliced* aligners such as STAR, or quantified without alignment; these are discussed in Chapter 7.

The output is stored in the **SAM** format, or its compressed binary equivalent **BAM** or reference-based **CRAM**, and processed with tools such as SAMtools (Li et al. 2009; Danecek et al. 2021). Before variant calling, reads are sorted, PCR duplicates are marked, and sometimes base qualities are recalibrated.

5.4.1.1 The SAM record

Each line of a SAM file describes one alignment of one read, in eleven tab-separated mandatory fields followed by optional tags (Li et al. 2009): the read name (QNAME), a bitwise **FLAG**, the reference sequence (RNAME), the 1-based leftmost position (POS), the mapping quality (**MAPQ**, see below), the **CIGAR** string, the position of the mate (RNEXT, PNEXT), the inferred insert size (TLEN), and the read sequence and base qualities (SEQ, QUAL). The FLAG packs yes/no properties into one integer, each bit meaning one thing: paired (1), properly paired (2), unmapped (4), mapped to the reverse strand (16), secondary alignment (256), PCR or optical duplicate (1024) and supplementary alignment (2048), the last used for the second part of a split read. A FLAG of 99 (1 + 2 + 32 + 64) is thus the first read of a properly mapped pair whose mate is on the reverse strand. The CIGAR string describes the alignment as runs of operations, for example `5S90M2I3M` for 5 soft-clipped bases, 90 aligned bases (match or mismatch), a 2-base insertion and 3 more aligned bases. Other operations are D (deletion), N (skipped reference, an intron in spliced RNA-seq alignments) and H (hard clipping). Optional tags give, for example, the edit distance (NM) and the read group (RG), which records the sample and sequencing run. Coordinate conventions differ between formats, 1-based in SAM and VCF but 0-based in BAM and BED (Section 2.4), which is a common source of off-by-one errors.

> **Practical quality control of sequencing data**
>
> Most errors in genome analyses come from the data, not from the algorithms. Before and after mapping, check:
>
> - **Base quality**: the distribution of Phred scores along the read (for example with FastQC); low-quality ends are trimmed.
> - **Adapter content**: adapter sequence appears when the insert is shorter than the read (Section 5.2) and must be trimmed (for example with cutadapt or fastp).
> - **Duplication**: PCR and optical duplicates inflate the apparent coverage and the support for errors; a high duplication rate indicates a library with few distinct molecules.
> - **Contamination**: reads from other organisms (bacteria, human DNA in non-human samples, adapter dimers) and cross-contamination between samples, detected by classifying reads against databases or from unexpected allele fractions.
> - **Coverage**: mean depth, and its uniformity, including GC bias and the fraction of the target covered at the depth needed for calling.
> - **Mapping**: the fraction of reads mapped, the distribution of mapping qualities, the fraction of properly paired reads and the insert-size distribution (for example with samtools stats).
>
> Tools such as MultiQC (Ewels et al. 2016) collect these statistics for many samples in one report, which makes outlier samples easy to see.

5.4.1.2 Mapping quality

A read from a repeat can align equally well to several locations. Each alignment is therefore given a **mapping quality**, a Phred-scaled probability that the reported location is wrong,

MAPQ $= -10 \log_{10} P(\text{wrong location})$ (Li et al. 2008). It is computed from the difference between the scores of the best and second-best hits: a read with two equally good hits gets a MAPQ of 0. Variant callers ignore or down-weight reads with low mapping quality, which means that variants in repeats and segmental duplications are often missed. Paired reads help, because one read in a unique region can anchor its mate in a repeat.

5.4.1.3 Reference bias and graph genomes

Mapping to a single linear reference genome introduces **reference bias**. Reads carrying the reference allele align slightly better than reads carrying a non-reference allele, so non-reference alleles are lost or under-counted, which distorts allele frequencies and allele-specific measurements. The effect is strongest around indels and in highly polymorphic or divergent regions, such as the HLA genes. Sequences absent from the reference, such as large insertions or population-specific sequence, cannot be mapped at all, and their reads may instead be mapped wrongly to a similar region elsewhere. Structural variants are therefore missed or mis-called.

Graph genomes, in which known variants are represented as alternative paths, reduce this bias (Garrison et al. 2018). This is no longer only a future idea. The draft human pangenome reference, a graph built from 47 phased diploid assemblies of individuals of diverse ancestry, was released in 2023 (Liao et al. 2023). Mapping short reads to it with graph-aware mappers in the vg toolkit (Garrison et al. 2018; Sirén et al. 2021) reduced small-variant errors by about a third and roughly doubled the number of structural variants detected per haplotype, compared with the linear GRCh38 reference (Liao et al. 2023). Linear references are still the default in most pipelines in 2026, but graph-based mapping and genotyping are in practical use and are expanding as more genomes are added.

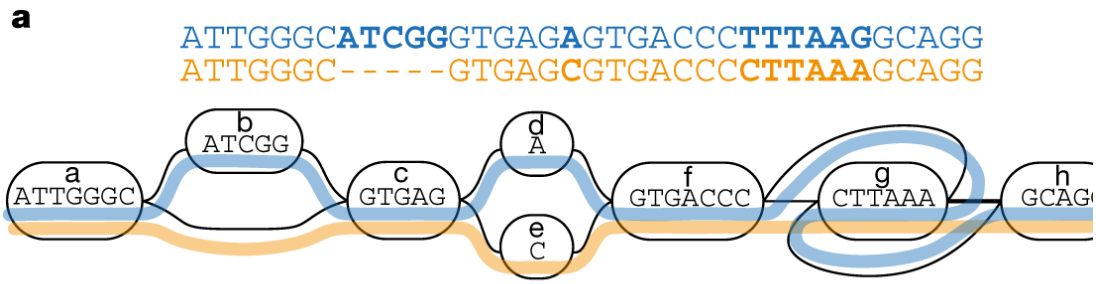


Figure 5.2: A sequence graph. Nodes hold DNA sequences and edges give the possible adjacencies; each haplotype is a walk through the graph (blue and orange), here differing by a deletion, a substitution and an inversion. Reproduced from Fig. 3a of Liao et al. (2023) (CC BY 4.0), as cropped in Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*.

A graph reference is not free of bias either: it represents only the haplotypes and variants used to build it, its coordinates are harder to report than positions on a linear chromosome, and adding many similar paths can make mapping more ambiguous. The ancestry of the genomes in the graph therefore matters, and graph and linear workflows should be compared on benchmarks stratified by population and region.

5.4.2 Variant calling

Genetic variants are classified by size:

- **SNPs** (single-nucleotide polymorphisms, or SNVs): single-base substitutions, the most common kind; two human genomes differ at about 4–5 million sites (The 1000 Genomes Project Consortium 2015);
- **indels**: small insertions and deletions, up to about 50 bases;
- **structural variants** (SVs): larger deletions, insertions, duplications (copy-number variants), inversions and translocations. They are fewer than SNPs but affect more bases in total.

Calls are stored in the **VCF** format, with the position, the alleles, the genotype of each sample and quality measures.

5.4.2.1 The VCF record

A VCF file has a header of `##` lines that define the reference, the filters and every annotation field, followed by one line per variant site (Danecek et al. 2011). The fixed columns are CHROM, POS (1-based), ID (for example a dbSNP `rs` number), REF, ALT (one or more alternative alleles, comma-separated), QUAL (a Phred-scaled probability that the site is not variant), FILTER (`PASS` or the names of failed filters) and INFO, a list of site-level annotations such as the total depth (DP) or allele frequency (AF). Then follow a FORMAT column and one column per sample, for example

```
chr1  1234567  .  ATCT  A  812  PASS  DP=64  GT:AD:GQ:PL  0/1:33,31:99:850,0,920
```

for a heterozygous 3-base deletion (an invented example). FORMAT names the per-sample fields: the genotype (GT, with 0 for REF and 1 for the first ALT allele; `/` for unphased and `|` for phased genotypes), the read depth for each allele (AD), the genotype quality (GQ) and the Phred-scaled genotype likelihoods (PL) described below. An indel is written with the preceding reference base included in both REF and ALT. The same indel in a repeat can be placed at several equivalent positions, so call sets must be **normalised** (left-aligned and split into one allele per line) before they are compared or looked up in databases.

5.4.2.2 Genotype likelihoods

Calling a SNP is a statistical problem, because sequencing errors are frequent compared with true variants. A heterozygous site is expected to show the alternative allele in about half of the reads, but with 10 reads, and an error rate of 1%, distinguishing a heterozygote from an error is not trivial. Modern callers use a probabilistic model (Marth et al. 1999; Li et al. 2008; Li 2011; Nielsen et al. 2011). For each possible genotype G at a site (for two alleles A and B in a diploid: AA, AB, BB) the **genotype likelihood** is the probability of the observed read bases D:

$$P(D \mid G = \{A_1, A_2\}) = \prod_i \left[\tfrac{1}{2} P(b_i \mid A_1) + \tfrac{1}{2} P(b_i \mid A_2) \right],$$

where b_i is the base in read i and $P(b_i \mid A) = 1 - e_i$ if $b_i = A$ and $e_i/3$ otherwise, with e_i the error probability given by the base quality (and, often, the mapping quality). The posterior probability of each genotype follows from Bayes' theorem with a prior, for example the expected heterozygosity, about 0.001 per site in humans. The genotype quality reported in the VCF file expresses the confidence in the called genotype.

It helps to keep three steps apart: the **read evidence** (the aligned bases with their base and mapping qualities), the **genotype likelihoods** computed from it, and the **genotype or variant call**, which also depends on the prior, on the assumed ploidy and on filters. A difference between the reads and the

reference is not automatically a biological variant. It may come from sequencing errors, from reads of a paralogous region mapped to the wrong place, from contamination with another sample or organism, or from a mixture of cells, as in a tumour or a pooled sample, where the diploid model with allele fractions of 0, ½ and 1 does not apply. When many individuals are sequenced, information can be shared between them through allele frequencies, which is particularly important at low coverage. Genotype likelihoods are also used directly in downstream population-genetic analyses, without calling genotypes at all.

5.4.2.3 GATK and haplotype-based calling

The **Genome Analysis Toolkit** (GATK) (McKenna et al. 2010; DePristo et al. 2011) defined the standard "best practices" pipeline for short-read data: duplicate marking, base-quality recalibration, variant calling, joint genotyping of many samples, and statistical filtering of the raw calls with a model trained on known variants (variant quality score recalibration). Its **HaplotypeCaller** does not call variants column by column. In regions that show evidence of variation, it locally re-assembles the reads into candidate haplotypes with a de Bruijn graph, aligns each read to each haplotype with a pair HMM, and then computes genotype likelihoods. This makes it much better at indels, which are often misaligned by the mapper. Similar tools include FreeBayes and the samtools/bcftools caller (Danecek et al. 2021). Somatic mutations in cancer, which may be present in only a small fraction of the cells, require specialised callers that compare tumour and normal samples (see below).

5.4.2.4 Deep-learning callers

DeepVariant (Poplin et al. 2018) recast variant calling as image classification. The reads around a candidate site are encoded as a multi-channel image (bases, base qualities, mapping qualities, strand), and a convolutional neural network, trained on genomes with known variants, predicts the genotype. DeepVariant outperformed the hand-crafted models, and it can be retrained for new technologies such as PacBio HiFi and ONT reads without changing the method. Variant callers are benchmarked against the high-confidence call sets of the Genome in a Bottle consortium, for which the true variants are known from many technologies (Zook et al. 2019).

5.4.2.5 Structural variants

Structural variants, including copy-number variants (CNVs), are biologically important: they cause many genetic diseases, are frequent driver events in cancer, and differ in more base pairs between two human genomes than SNPs do. They cannot be found by examining one alignment column at a time, as small variants are, because reads that cross an SV breakpoint often fail to align or align only in part. They are therefore detected from different kinds of evidence, four kinds of signal in short-read data (Mahmoud et al. 2019):

- **read depth**: deletions and duplications change the number of reads mapped to a region;
- **discordant read pairs**: pairs that map too far apart, too close or in the wrong orientation indicate deletions, insertions or inversions;
- **split reads**: reads whose two parts map to different locations show the exact breakpoint;
- **assembly**: local or whole-genome assembly and comparison with the reference.

Each signal sees only some kinds of SV. Read depth detects copy-number changes but not balanced events, such as inversions and translocations, and gives only approximate breakpoints. Discordant pairs and split reads locate breakpoints, but cannot reveal the content of insertions longer than the fragment, and fail when the breakpoint lies in a repeat. Short-read SV callers therefore combine several signals.

With short reads, SV calling is much less accurate than SNP calling, and many SVs in repeats are missed entirely. Long reads span most SVs, and long-read callers such as Sniffles (Sedlazeck et al. 2018) and comparisons of haplotype-resolved assemblies have shown that each human genome carries more than 20,000 SVs, most of which were invisible to short reads.

5.4.2.6 Phasing

A diploid genotype such as A/G at one site and C/T at another does not say which alleles lie on the same chromosome. **Phasing** assigns alleles to the two *haplotypes*. This matters biologically: for a recessive disorder, two loss-of-function mutations in the same gene cause disease if they are on different copies (*compound heterozygosity*), but not if they are on the same copy, since the other copy is then intact. Haplotypes are also needed for allele-specific expression, for imputation of untyped variants, and for studies of recombination and selection. There are three main approaches:

- **read-backed phasing**: reads or read pairs that cover two or more heterozygous sites show directly which alleles occur together. Long reads give phased blocks of many kilobases to megabases (Patterson et al. 2015);
- **pedigree (trio) phasing**: the alleles inherited from each parent are determined from the parents' genotypes;
- **statistical (population) phasing**: haplotypes are inferred from patterns of linkage disequilibrium in a large reference panel of haplotypes, using hidden Markov models of how a new haplotype is a mosaic of the haplotypes in the panel. This works well for common variants but not for rare or private ones (Delaneau et al. 2019).

Read-backed phasing can be formulated as a combinatorial problem. In a matrix with one row per read and one column per heterozygous site, each entry gives the allele the read shows. With error-free reads the rows split cleanly into two groups, one per haplotype. With errors, the **minimum error correction** (MEC) problem asks for the partition that requires the fewest changes to the matrix, with each change weighted by the base quality. MEC is NP-hard, but WhatsHap solves it exactly in time exponential only in the read coverage, not in the number of reads or sites, which makes it practical for long reads at moderate depth (Patterson et al. 2015).

Phasing produces *phase blocks*, within which the relative phase is known, and its errors are measured as *switch errors* between adjacent sites. Long blocks with many switch errors can be worse than short accurate ones, so block length and switch error rate should be reported together. Haplotype-resolved assembly (Section 5.3) gives phasing over whole chromosomes.

5.4.3 Targeted and exome sequencing

Much clinical and cancer sequencing does not cover the whole genome. In **exome sequencing**, the protein-coding exons, about 1–2% of the genome, are enriched by **hybrid capture**: the DNA library is hybridised to biotinylated oligonucleotide baits complementary to the targets, which are pulled down with streptavidin beads before sequencing (Gnirke et al. 2009). Smaller **gene panels** use the same principle, or amplify the targets by PCR (amplicon sequencing). Targeting makes high depth affordable, which is needed for rare variants in tumours, but brings its own biases:

- **uneven coverage**: capture efficiency depends on GC content, bait design and the mappability of the target, so some exons are covered at hundreds of times the depth of others, and some, such as GC-rich first exons, hardly at all. The fraction of the target covered at, say, 20-fold is a more useful summary than the mean depth;
- **off-target reads**: a substantial fraction of reads come from outside the baits. They are wasted for small-variant calling, but give low-coverage information on copy number genome-wide;

- **reference bias in capture**: baits designed from the reference capture divergent alleles and large indels less efficiently;
- **duplicates**: with little input DNA, many reads come from the same molecule, and unique molecular identifiers are used to tell them apart.

Variants outside the targets, including most regulatory and structural variants, are invisible, and copy-number calling from exome data is noisier than from genomes, because the depth varies so much between targets. As sequencing costs fall, genome sequencing is replacing exome sequencing in many clinical settings.

5.4.4 Copy-number analysis

Copy-number variants (CNVs), deletions and duplications from about a kilobase to whole chromosome arms, are detected from sequencing **read depth**. The genome (or the targets) is divided into bins, reads are counted in each, and the counts are corrected for GC content and mappability and normalised against a reference, a matched normal sample or a panel of normal samples, which removes biases shared by all samples. The resulting log ratios are then **segmented** into regions of constant copy number, with change-point methods such as circular binary segmentation or with hidden Markov models whose states are copy numbers. Tools such as CNVkit use both on-target and off-target reads from targeted sequencing (Talevich et al. 2016).

Before sequencing became cheap, CNVs were measured with SNP arrays, which remain common in clinical genetics and in biobanks. Each SNP gives two quantities (Peiffer et al. 2006). The **log R ratio** (LRR) is the log of the total intensity of both alleles relative to that expected for two copies, and falls with deletions and rises with duplications. The **B-allele frequency** (BAF) is the fraction of the signal from one allele, and takes the values 0, ½ and 1 for the three genotypes of a normal diploid region. A heterozygous deletion removes the ½ band; a duplication splits it into ⅓ and ⅔; and **copy-neutral loss of heterozygosity**, in which a region is inherited from one parent twice or a tumour has lost one homologue and duplicated the other, removes the ½ band without changing the LRR. The same two signals are computed from read depth and allele fractions in sequencing data, and are the basis of copy-number analysis in cancer.

5.4.5 Somatic variants in cancer

A tumour sample differs from a germline sample in three ways: it is a mixture of tumour and normal cells, it is often aneuploid, and the tumour cells themselves are genetically heterogeneous. The diploid genotype model of the previous sections does not apply, and **somatic variant calling** compares a tumour with a **matched normal** sample from the same person, usually blood, so that germline variants can be removed. Callers such as Mutect (Cibulskis et al. 2013) ask whether the alternative allele in the tumour is better explained by a somatic mutation present at some unknown fraction than by sequencing error, and whether it is absent from the normal.

The **variant allele fraction** (VAF), the fraction of reads that carry the variant, depends on the tumour **purity** p, the copy number C_T at the locus in the tumour, the number of copies carrying the mutation m, and the fraction of tumour cells with the mutation. For a clonal mutation,

$$\mathrm{VAF} = \frac{p\,m}{p\,C_T + 2(1-p)}.$$

A heterozygous clonal mutation in a diploid region of a tumour with 40% purity is thus seen in only 20% of reads, and subclonal mutations at even lower fractions, which requires high depth. Purity and ploidy are estimated jointly from the copy-number profile and the allele fractions, for example by ABSOLUTE

(Carter et al. 2012); together they turn VAFs into the fraction of cancer cells carrying each mutation, which is the starting point for reconstructing subclones and tumour evolution (Section 7.7).

Somatic calls are dominated by artefacts unless they are filtered. A **panel of normals**, a set of unrelated normal samples processed in the same way, removes recurrent technical artefacts and common germline variants missed in the matched normal. Other artefacts come from sample preparation, such as C>T changes from formalin fixation and G>T changes from oxidative damage, and many are strand- or read-position-specific. Normal tissue is also not truly normal: blood can carry clonal haematopoiesis mutations, and a normal sample contaminated with tumour cells hides real somatic variants.

5.4.5.1 Drivers, passengers and mutational signatures

A typical adult tumour carries thousands of somatic point mutations, but only a handful are **drivers** that confer a growth advantage; the rest are **passengers** that happened to be present in the clone that expanded (Stratton et al. 2009; Vogelstein et al. 2013). Drivers are identified statistically, as genes mutated more often than expected from the background mutation rate. This rate varies along the genome with expression and replication timing, and failing to model it produced many false cancer genes (Lawrence et al. 2013). An alternative is to apply the ratio of non-synonymous to synonymous mutations, dN/dS (Section 3.6), to somatic mutations, which shows positive selection in drivers and nearly no negative selection in most of the tumour genome (Martincorena et al. 2017). Curated lists of driver genes are kept in the Cancer Gene Census of COSMIC (Section 2.9).

Passengers are not useless: they record the mutational processes that acted on the tumour. Each single-base substitution is classified into one of 96 types, six substitution classes (C>A, C>G, C>T, T>A, T>C, T>G, with the pyrimidine of the base pair as reference) times sixteen combinations of flanking bases. The mutation counts of many tumours form a matrix that is decomposed by non-negative matrix factorisation into **mutational signatures** and their activity in each tumour (Alexandrov et al. 2013). Several signatures have known causes: C>A changes from tobacco smoke in lung cancer, C>T changes at dipyrimidines from ultraviolet light in melanoma, a clock-like C>T signature at CpG sites from deamination of methylated cytosine, and signatures of APOBEC cytidine deaminases and of defective DNA repair. The current COSMIC catalogue, extended to doublet substitutions and indels from thousands of whole genomes (Alexandrov et al. 2020), is used to infer exposures and, for example, defects in homologous recombination that predict response to PARP inhibitors.

See also: further explanation in Section D.5.3; related Wikipedia articles in Section E.5.3.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Sequence alignment, SAM (file format), SNV calling from NGS data, Variant Call Format and Structural variation (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Sequencing, Mapping, Assembly, and Variation” (CC BY 4.0), used with permission.

5.5 Regulatory and epigenomic sequencing data

Sequencing is used not only to read genomes, but also to measure where proteins bind DNA and RNA, which parts of the genome are accessible, and which bases are chemically modified. These assays turn a biochemical property into a set of sequenced DNA fragments, which are mapped to the reference with the same tools as genomic reads (Section 5.4). The computational problem is then to find regions where

reads are enriched, or bases where the sequence has been altered, against a background that is far from uniform. The regulatory elements found in this way are described in Section 6.9.

5.5.1 ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) maps the binding sites of a transcription factor, or the positions of a histone modification, across the genome (Johnson et al. 2007). Proteins are cross-linked to DNA, the chromatin is fragmented, fragments bound by the protein are pulled down with an antibody, and the DNA is sequenced. Reads pile up around the binding sites.

The background is not uniform: open chromatin fragments more easily, some regions map better than others, and amplified regions in the sample give more reads. Every experiment therefore needs a **control**, usually the **input** DNA taken before immunoprecipitation. **Peak callers** such as MACS (Zhang et al. 2008) find regions where the ChIP signal exceeds the control. MACS uses the fact that reads from the two strands form two peaks shifted by about one fragment length on either side of the binding site, shifts the reads by half this distance to centre them, and tests for enrichment against a *local* Poisson background estimated from the control in the surrounding kilobases. Transcription factors give sharp, narrow peaks; histone marks such as H3K27me3 cover broad domains and need different settings.

Biological replicates rarely agree perfectly. The **irreproducible discovery rate** (IDR) ranks peaks in two replicates and estimates which part of the ranked list is reproducible, much as a false discovery rate estimates which part is significant (Li et al. 2011). The ENCODE guidelines use IDR, the **fraction of reads in peaks** and other measures to judge the quality of experiments and antibodies (Landt et al. 2012). Finally, the sequences under the strongest peaks are searched for enriched motifs, both to confirm that the experiment worked (the factor's own motif should be at the centre of the peaks) and to find co-binding partners (Section 1.6) (Heinz et al. 2010). A peak shows where the protein is cross-linked, directly or indirectly, not that the binding regulates a gene.

5.5.2 Chromatin accessibility: DNase-seq and ATAC-seq

Active regulatory elements are depleted of nucleosomes, which makes them accessible to enzymes. **DNase-seq** digests chromatin lightly with DNase I and sequences the cut sites (Boyle et al. 2008). **ATAC-seq** uses the Tn5 transposase, which inserts sequencing adapters preferentially into open chromatin, and works with tens of thousands of cells or fewer, and also in single cells (Buenrostro et al. 2013). Peak calling resembles that of ChIP-seq, but no antibody is needed and one experiment finds all accessible elements at once. The fragment lengths reveal nucleosome positions, and short protected *footprints* within peaks can indicate bound transcription factors, although enzyme sequence bias complicates footprinting. Typical quality checks are the enrichment of reads at transcription start sites and the fraction of reads from mitochondrial DNA, which is accessible and often dominates poor libraries.

5.5.3 DNA methylation

In mammals, DNA methylation occurs mainly as 5-methylcytosine in CpG dinucleotides and is associated with silenced promoters and repeats (Section 6.7). In **bisulfite sequencing**, sodium bisulfite converts unmethylated cytosines to uracil, read as thymine after PCR, while methylated cytosines are protected (Frommer et al. 1992). Methylation is thus turned into a C/T difference between reads and reference.

This complicates mapping. Converted reads contain almost no C (or, from the other strand, no G) and match the reference poorly. Aligners such as Bismark convert both reads and genome *in silico* (C to T, and G to A for the opposite strand), map to the converted genomes, and then compare the original read with the reference (Krueger and Andrews 2011). The **methylation level** of a cytosine is the

fraction of reads that show C rather than T at that position. Whole-genome bisulfite sequencing gave the first base-resolution methylomes, which revealed widespread non-CpG methylation in embryonic stem cells (Lister et al. 2009). Pitfalls include incomplete conversion, which mimics methylation, C-to-T SNPs, which mimic unmethylated sites, and the inability of bisulfite to distinguish 5-methylcytosine from 5-hydroxymethylcytosine. Bisulfite also degrades DNA, and enzymatic conversion methods are a gentler alternative. Cheaper alternatives are reduced-representation bisulfite sequencing, which covers CpG-rich regions, and methylation arrays.

Long reads measure methylation without conversion (Section 5.2). In nanopore data a methylated base changes the current signal, and methylation is called by comparing the signal with models of methylated and unmethylated *k*-mers (Simpson et al. 2017), now usually with neural networks built into the basecaller. PacBio detects methylation from the kinetics of the polymerase. Because each read comes from one molecule, long reads also give methylation on each haplotype, for example at imprinted genes.

5.5.4 RNA–protein interactions: CLIP-seq

CLIP-seq (cross-linking and immunoprecipitation) maps the binding sites of RNA-binding proteins. Cells are irradiated with UV light, which cross-links proteins to the RNA they touch; the protein is immunoprecipitated, the RNA trimmed and sequenced (Licatalosi et al. 2008). The cross-linked amino acid leaves a trace in the reads, a mutation or a truncation of the reverse transcription, which locates binding sites with nucleotide resolution. As in ChIP-seq, a control is essential, and in eCLIP a size-matched input sample serves as background (Van Nostrand et al. 2016). Enrichment must be judged against the abundance of the RNA itself, since highly expressed transcripts give reads in any experiment.

See also: further explanation in Section D.5.6; related Wikipedia articles in Section E.5.6.

> **i** Sources
>
> Written for this book from the primary literature cited in the text.

5.6 Gene prediction

A genome sequence becomes useful only when its genes have been found. **Gene prediction** (gene finding) identifies the regions that encode proteins or functional RNAs, and **genome annotation** adds the gene structures, their products and their functions (Yandell and Ence 2012). This section concentrates on protein-coding genes. The prediction of non-coding RNA genes is discussed in Section 1.7, and functional annotation in Chapter 10.

5.6.1 Signals and content

Gene finders combine two kinds of information:

- **Signals**: short sequence motifs at defined positions, such as start and stop codons, splice sites (almost all introns start with GT and end with AG), ribosome-binding sites and promoters. They are modelled with position weight matrices or similar models (Section 1.6).

- **Content**: statistical differences between coding and non-coding sequence. Because the genetic code is read in triplets and codon usage is biased, coding sequence has a characteristic three-base periodicity and hexamer composition. Fickett showed in 1982 that such statistics can recognise coding regions (Fickett 1982).

The **open reading frame** (ORF), a stretch without stop codons in one reading frame, is the simplest content feature. In random sequence a stop codon appears on average every 21 codons, so ORFs longer than a few hundred codons are unlikely to occur by chance.

5.6.2 Prokaryotic gene finding

Prokaryotic genomes are dense, with about 85–90% of the sequence coding, and genes are not interrupted by introns. Gene finding therefore reduces to deciding which long ORFs are real genes and where they start. Statistical methods use **Markov models** of coding sequence: the probability of each base depends on the preceding k bases and on its position in the codon (a *three-periodic*, or inhomogeneous, Markov chain). **GeneMark** (Borodovsky and McIninch 1993) used fifth-order Markov chains for each of the six reading frames, and **GeneMark.hmm** (Lukashin and Borodovsky 1998) embedded them in a hidden Markov model of the whole genome, which also models ribosome-binding sites and gene overlaps. **Glimmer** (Delcher 1999) uses *interpolated* Markov models, which combine models of different orders depending on how much training data are available for each context. **Prodigal** (Hyatt et al. 2010) improved the prediction of translation start sites and is widely used in pipelines, including for metagenomes.

A simple way to score an ORF x is the log-likelihood ratio of a coding and a non-coding model,

$$S(x) = \log \frac{P(x \mid \text{coding})}{P(x \mid \text{non-coding})} = \sum_{i} \log \frac{P(x_i \mid x_{i-k} \dots x_{i-1}, \phi_i)}{P(x_i \mid x_{i-k} \dots x_{i-1})},$$

where ϕ_i is the codon position of base i. Scores for start codons, ribosome-binding sites, length and overlaps are added, and a dynamic programming pass along the genome selects a compatible set of genes. The score is relative to the model: a real gene with unusual composition can score poorly.

These programs are trained on the genome itself (*self-training*), since long ORFs are almost certainly genes. Glimmer, for example, first collects long non-overlapping ORFs, trains its interpolated Markov model on them, and then scores all candidate ORFs. Training on long ORFs biases the model against short genes, and the minimum gene length is a trade-off between sensitivity and false positives. A pile-up of predictions at exactly the minimum length is a sign that this threshold, rather than the biology, shapes the result. They find well over 95% of genes. The remaining errors are mostly in start sites, very short genes, genes acquired by horizontal transfer with an unusual composition (Section 3.7), and genes that are pseudogenised. Production pipelines, such as the NCBI prokaryotic genome annotation pipeline (Tatusova et al. 2016), combine ab initio predictions with protein homology.

5.6.3 Eukaryotic gene finding

Eukaryotic gene prediction is much harder. Genes are split into exons separated by introns that can be hundreds of kilobases long, coding sequence makes up only about 1–2% of the human genome, and most genes have several alternatively spliced transcripts (Section 3.5). A gene finder must predict not only where a gene is, but the exact boundaries of every exon. Eukaryotic gene prediction is therefore not simply ORF detection plus splice-site recognition. A complete gene model also includes the untranslated regions (UTRs) and the alternative isoforms of each gene, and a genome also contains many non-coding RNA genes. Classical ab initio gene finders predict one coding structure per locus and no UTRs, so these

features must come from transcript evidence, which is itself often incomplete, for example when cDNA or assembled transcripts lack their 5' ends.

Generalised hidden Markov models. The breakthrough came with **GENSCAN** (Burge and Karlin 1997), which models the structure of a whole gene with a *generalised* (semi-Markov) hidden Markov model. Each state represents a whole segment, such as an initial, internal or terminal exon, an intron in one of three phases, an intergenic region or a promoter, and emits a sequence of variable length with an explicit length distribution. Exon states use codon-position-specific Markov models of content, and the transitions between states are scored with signal models for splice sites and start and stop codons. The most probable parse of the sequence into gene structures is found with a Viterbi-like dynamic programming algorithm (Section 1.6). Explicit length distributions are important because real exons and introns have length distributions far from the geometric distributions of ordinary HMM states. Earlier, GeneID had shown how candidate exons could be scored and assembled into genes by dynamic programming (Guigó et al. 1992). Successors include **AUGUSTUS** (Stanke and Waack 2003), with an improved intron model and the ability to integrate external evidence, and **SNAP**, designed to be trained easily on new genomes (Korf 2004).

Accuracy. Gene finders are evaluated at the nucleotide, exon and gene levels, with sensitivity (the fraction of true features predicted) and specificity (the fraction of predictions that are true) (Burset and Guigó 1996). Note that "specificity" in the gene-finding literature means what is elsewhere called *precision*, not the true-negative rate (4). At the nucleotide level, accuracy is high. At the gene level, where every exon must be exactly right, purely ab initio predictions of human genes are correct for only a minority of genes. Small exons, untranslated regions, alternative splicing and genes within genes are the hardest parts.

5.6.4 Evidence-based annotation

Since ab initio prediction alone is not accurate enough for eukaryotes, genome annotation pipelines combine it with external **evidence**:

- **Transcript evidence**: cDNA and EST sequences, and now RNA-seq reads aligned with spliced aligners or assembled into transcripts, and full-length transcripts from long-read sequencing (Iso-Seq or direct RNA), which show exon–intron structures directly.
- **Protein evidence**: proteins from related species aligned to the genome with *spliced* protein-to-genome alignment, which identifies exons by homology even when the genes are not expressed in the sampled tissues.
- **Comparative evidence**: conservation between related genomes, since coding exons evolve more slowly and show the three-periodic pattern of synonymous changes.

Before gene prediction, **repeats** are identified and masked, since transposable elements contain ORFs that would otherwise be predicted as genes. Pipelines such as MAKER (Holt and Yandell 2011) and BRAKER (Brůna et al. 2021) align the evidence, use it to train and guide gene finders such as AUGUSTUS and GeneMark, and combine the results into consensus gene models. BRAKER3 combines RNA-seq and protein evidence in one pipeline (Gabriel et al. 2024). MAKER also reports for each gene model an *annotation edit distance* (AED), from 0 for complete agreement with the aligned evidence to 1 for no agreement (Holt and Yandell 2011). AED measures agreement with the evidence, not correctness: a real gene that is not expressed in the sampled tissues gets a poor score. Figure 5.3 summarises such a pipeline.

The evidence sources are not fully independent. A protein in a database may itself be a computational prediction, so using it to support another prediction can propagate an error from genome to genome. A predicted gene is well supported when RNA-seq reads confirm its splice junctions, a protein from a related species aligns along its whole length, and its domains occur in the expected order. A protein

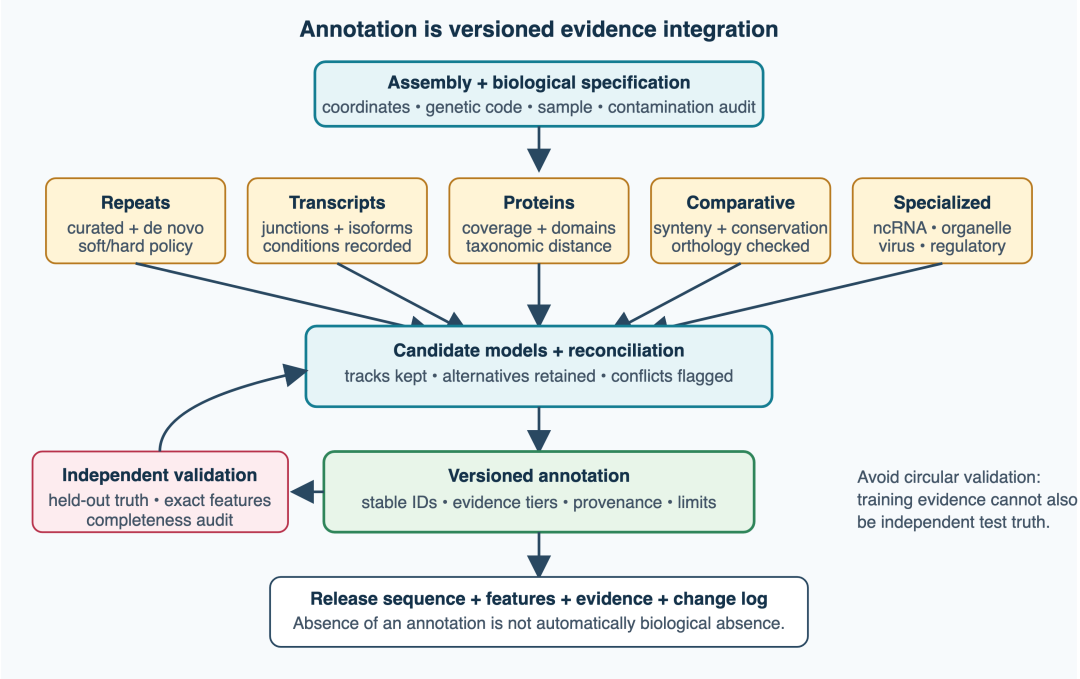


Figure 5.3: Genome annotation integrates several kinds of evidence (repeats, transcripts, proteins, conservation, specialised predictors for non-coding RNAs and other features) into candidate gene models, which are reconciled and released as versioned annotations. Evidence used to train the gene finders cannot also serve as independent test data. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

match that covers only one domain, an internal stop codon or a frameshift suggests an incomplete or wrong gene model. Equally, the absence of a gene from an annotation does not show that the gene is absent from the genome, since it may lie in a gap of the assembly or have been missed by the pipeline. The Ensembl, RefSeq and GENCODE annotations of the human genome are curated with such pipelines and manual review. **Deep learning** has entered this field too: Helixer predicts gene structures directly from the genome sequence with a neural network and needs no species-specific training (Stiehler et al. 2020).

Even for the human genome, the number of protein-coding genes was debated for years: estimates before the genome project ranged up to 100,000, the first drafts predicted about 30,000–40,000, and the number in current annotations (2026) is about 20,000 (Chapter 6).

See also: further explanation in Section D.5.4; related Wikipedia articles in Section E.5.4.

i Sources

Parts of this section are adapted from the Wikipedia articles Gene prediction, Open reading frame, GENSCAN, GeneMark and Genome project (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapters “Genome Architecture, Representation, and Annotation” and “Genome Architecture and Gene Prediction” (CC BY 4.0), used with permission.

5.7 Metagenomics

Metagenomics is the study of genetic material recovered directly from an environment, without first isolating and culturing the organisms. Most microbes cannot be grown in the laboratory, so metagenomics has changed our view of microbial diversity, from the oceans and soil to the human gut, and of the tree of life itself (Section 4.5). It also raises its own algorithmic problems, because the data are a mixture of many genomes in unknown proportions (Quince et al. 2017).

5.7.1 History

In the 1980s Norman Pace and colleagues began to sequence ribosomal RNA genes amplified directly from environmental samples, which showed that microbial diversity was vastly greater than that of cultured species (Pace 1997). The term *metagenome* was introduced in 1998 for the collective genomes of the microbes in soil (Handelsman et al. 1998). The first large shotgun metagenomes followed in 2004: nearly complete genomes were reconstructed from a simple community in acid mine drainage (Tyson et al. 2004), and Venter's Sargasso Sea project reported more than a million new genes (Venter et al. 2004). Large projects then characterised the human gut (Qin et al. 2010) and the human microbiome at many body sites (The Human Microbiome Project Consortium 2012), and the Tara Oceans expedition the global ocean microbiome (Sunagawa et al. 2015). Catalogues of hundreds of thousands of genomes assembled from metagenomes now describe most of the species in the human gut (Almeida et al. 2019; Pasolli et al. 2019) and many environments on Earth (Nayfach et al. 2020).

5.7.2 Amplicon sequencing

The simplest approach is **amplicon** (marker-gene) sequencing: a variable region of the 16S rRNA gene (for bacteria and archaea), or of the 18S rRNA or ITS region (for eukaryotes and fungi), is amplified with universal primers and sequenced. It is cheap and sensitive, and is still the most common way to describe community composition.

Reads were traditionally clustered into **operational taxonomic units** (OTUs) at 97% identity, a rough proxy for species. Since clustering merges real variants with sequencing errors, modern methods instead infer the exact **amplicon sequence variants** (ASVs) in the sample, using a model of the sequencing errors to decide which sequences are real (Callahan et al. 2016). ASVs are reproducible across studies. Pipelines such as QIIME 2 (Bolyen et al. 2019) assign taxonomy by comparison with reference rRNA databases and compute diversity measures: *alpha diversity* within a sample and *beta diversity* between samples, using distances that may take phylogeny into account.

Amplicon data have well-known limitations. PCR primers do not amplify all taxa equally, the number of rRNA operons differs between genomes, a short region of 16S often cannot distinguish species or strains, and the data tell nothing about the functions encoded in the genomes.

5.7.3 Shotgun metagenomics

In **shotgun metagenomics**, all DNA in the sample is sequenced. This gives access to all organisms (including viruses and eukaryotes) and to their genes, but the analysis is much harder (Quince et al. 2017). The main tasks are:

- **Taxonomic profiling**: which organisms are present, and in what proportions. Read classifiers such as **Kraken** (Wood and Salzberg 2014; Wood et al. 2019) assign each read to the lowest common ancestor of the genomes that share its *k*-mers (Section 1.5). Marker-gene profilers such as **MetaPhlAn** (Blanco-Míguez et al. 2023) map reads to a set of clade-specific genes, which gives more accurate abundances. Profiling is limited by the reference database: organisms without sequenced relatives are invisible or misclassified.
- **Functional profiling**: which genes and pathways are present. Reads or predicted genes are compared with protein databases, usually with fast translated search such as DIAMOND or MMseqs2 (Section 1.4), and summarised by gene families and pathways (Chapter 10). Gene catalogues, such as the 3.3 million non-redundant genes of the first MetaHIT study of the gut (Qin et al. 2010), serve as references.
- **Assembly**: reconstructing genomes without a reference. Metagenome assembly is harder than single-genome assembly (Section 5.3), because coverage varies by orders of magnitude between species, closely related strains create bubbles and tangles in the graph, and repeats are shared between genomes. Assemblers such as metaSPAdes (Nurk et al. 2017) and MEGAHIT (Li et al. 2015) are adapted to these problems.
- **Binning**: grouping contigs that come from the same genome, described below.

Profiling and assembly answer different questions. A profiler estimates which *known* taxa are present and in what proportions, by comparing reads with reference genomes or marker genes, and does not reconstruct genomes. Assembly and binning reconstruct genome sequences, including those of organisms with no sequenced relatives, but only for organisms with enough coverage, and they say less about the rare members of the community.

Genes are predicted on the contigs, mostly with prokaryotic gene finders such as Prodigal (Section 5.6). The proteins found in metagenomes now far outnumber those from cultured organisms, and are a main source of sequences for protein families, structure prediction and protein language models (Section 1.10).

5.7.4 Binning and metagenome-assembled genomes

Binning groups contigs into putative genomes, **metagenome-assembled genomes** (MAGs). It uses two kinds of signal:

- **sequence composition**: different genomes have different GC content and oligonucleotide (typically tetranucleotide) frequencies, while contigs from the same genome are similar;
- **coverage**: contigs from the same genome have similar coverage within a sample, and, more powerfully, their coverage varies in the same way across many samples (*co-abundance*).

Tools such as MetaBAT 2 (Kang et al. 2019) combine the two signals to cluster contigs. The quality of each MAG is then estimated from sets of single-copy marker genes: the fraction of expected markers present gives the *completeness*, and the fraction present in multiple copies the *contamination* (Parks et al. 2015). By community standards, a high-quality draft MAG is more than 90% complete with less than 5% contamination and contains the rRNA genes and most tRNAs (Bowers et al. 2017). A MAG is not equivalent to the genome of an isolate. It is typically a consensus of closely related strains in the sample, it often lacks mobile elements, plasmids and rRNA operons, which are hard to assemble and bin, and it may contain contigs from other organisms that the marker-gene estimates do not detect. Long-read sequencing increasingly produces complete, circular genomes directly from metagenomes.

MAGs have changed microbial phylogeny: much of the diversity in current trees of life, including the candidate phyla radiation and the Asgard archaea, is known only from metagenomes (Section 4.5). The Genome Taxonomy Database includes MAGs in its classification (Parks et al. 2018).

5.7.5 Benchmarking

As in other fields, community challenges have been used to compare methods. The Critical Assessment of Metagenome Interpretation (CAMI) provides realistic simulated and mock communities with known composition, and has evaluated assemblers, binners and profilers (Sczyrba et al. 2017; Meyer et al. 2022). The results show that methods work well for species with a single strain at high coverage, and poorly for closely related strains and for low-abundance organisms.

See also: further explanation in Section D.5.5; related Wikipedia articles in Section E.5.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Metagenomics, 16S ribosomal RNA, Amplicon sequence variant, Binning (metagenomics) and Metagenome-assembled genome (CC BY-SA 4.0, retrieved 3 October 2026).

5.8 Human genetics: finding disease genes and variants

Variant calling (Section 5.4) produces a list of differences between a genome and the reference. Human genetics asks which of these differences affect a trait or cause a disease. Two very different strategies are used. For common, complex diseases, millions of variants are tested for statistical association in large populations. For rare, Mendelian diseases, the genome or exome of a patient, often with the parents, is searched for one or two variants that explain the disease. Both depend on how variants are inherited together along the chromosome.

5.8.1 Linkage disequilibrium and haplotype structure

Alleles at nearby sites are not inherited independently. A new mutation arises on one particular haplotype, and only recombination, a few events per chromosome per generation, separates it from its neighbours. The non-random association of alleles at different sites is called **linkage disequilibrium** (LD). For two biallelic sites with alleles A/a and B/b, it is measured by

$$D = p_{AB} - p_A p_B,$$

the difference between the observed frequency of the AB haplotype and the frequency expected if the sites were independent. Because the range of D depends on the allele frequencies, two normalised measures are used. $D' = D/D_{\max}$ is 1 when at most three of the four haplotypes are present, which shows that no recombination (or recurrent mutation) has been observed between the sites. The squared correlation

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

is 1 only when the two sites carry the same information, so that one predicts the other perfectly. r^2 is what matters for association studies: a variant in LD r^2 with a causal variant captures its effect with a sample size about $1/r^2$ times larger.

LD decays with distance, but not smoothly. In humans most recombination occurs in **hotspots** of one or two kilobases, positioned largely by the DNA-binding protein PRDM9 (Myers et al. 2005). Between hotspots, the genome forms **haplotype blocks** of tens of kilobases, within which only a few common

haplotypes exist (Gabriel et al. 2002). The International HapMap Project mapped these patterns in several populations ("A Haplotype Map of the Human Genome" 2005). A few **tag SNPs** chosen from each block capture most of the common variation (Carlson et al. 2004), which is why a genotyping array with half a million SNPs can survey the common variation of a whole genome. LD extends further in populations that have passed through bottlenecks, such as Europeans and East Asians, than in African populations, which affects both the design and the transferability of genetic studies.

5.8.2 Genetic architecture of disease

In **Mendelian** (monogenic) diseases, such as cystic fibrosis or Huntington's disease, a variant in a single gene is sufficient to cause the disease. Such variants are rare, because selection keeps them at low frequency, and have large effects. **Complex** diseases and traits, such as type 2 diabetes, schizophrenia or height, are influenced by many variants of small effect together with the environment. In between lie variants with incomplete **penetrance**, the probability that a carrier develops the disease, such as *BRCA1* mutations in breast cancer. Effect size and allele frequency are therefore inversely related: common variants have small effects, and large effects are found mostly in rare variants. The methods differ accordingly: association studies find common variants of small effect, and sequencing of families finds rare variants of large effect.

Heritability is the fraction of the variance of a trait in a population that is explained by genetic differences. Twin studies estimate it to be about 80% for height, yet the first GWAS hits together explained only a few per cent, the problem of **missing heritability** (Manolio et al. 2009). Much of it was later found to be hidden in very many common variants, each too weak to reach significance: fitting all SNPs together explained about half of the heritability of height (Yang et al. 2010). The rest is thought to lie in rare variants, poorly tagged variants and, possibly, overestimated twin heritability.

5.8.3 Linkage analysis

Before association studies, disease genes were mapped by **linkage analysis** in families. Markers spread over the genome are genotyped in affected and unaffected relatives, and one tests whether a marker is co-inherited with the disease more often than expected for an unlinked locus. The evidence is summarised as a **LOD score**, the base-10 logarithm of the ratio between the likelihood of the data at a recombination fraction $\theta < 0.5$ and at $\theta = 0.5$ (no linkage). A LOD score of 3, odds of 1000 to 1, was the classical threshold, and thresholds for genome-wide scans were later formalised (Lander and Kruglyak 1995). Linkage analysis found the genes for hundreds of Mendelian diseases, such as *CFTR* and *HTT*, but it has low power for complex diseases and low resolution, since a family has few recombinations and a linked region can contain hundreds of genes.

5.8.4 Genome-wide association studies

A **genome-wide association study** (GWAS) tests each of millions of variants for association with a trait in unrelated individuals, either in a **case–control** design, comparing allele frequencies between patients and controls with logistic regression, or for a **quantitative trait** with linear regression of the trait on the number of alternative alleles (0, 1 or 2). The Wellcome Trust Case Control Consortium study of 14,000 cases of seven diseases and 3,000 shared controls established the design in 2007 (Burton et al. 2007), and tens of thousands of associations with thousands of traits have since been reported (Visscher et al. 2017; Uffelmann et al. 2021) and collected in the GWAS Catalog (Section 2.9).

Most GWAS genotype samples with **SNP arrays** (Section 7.2) rather than by sequencing. The missing variants are then **imputed**: each individual's haplotypes are modelled as a mosaic of haplotypes in a

sequenced **reference panel**, such as 1000 Genomes or TOPMed, with the same kind of hidden Markov model used for statistical phasing (Section 5.4), which predicts genotypes at millions of untyped sites (Marchini and Howie 2010). Imputation works well for common variants in populations represented in the panel, and poorly for rare variants.

Multiple testing. The conventional genome-wide significance threshold is $P < 5 \times 10^{-8}$. It is a Bonferroni correction of 0.05 for about one million independent tests, the effective number of independent common variants in the genome of European-ancestry populations, estimated from LD (Pe'er et al. 2008). It is not tied to the number of variants actually tested: imputing more variants adds tests, but few independent ones. Because the threshold is strict, true associations of small effect need very large samples, and modern GWAS include hundreds of thousands to millions of individuals. Associations must be **replicated** in an independent sample, and the first reported effect size is usually inflated (the *winner's curse*).

Population stratification. If cases and controls differ in ancestry, every variant whose frequency differs between the ancestral groups will appear associated with the disease. **EIGENSTRAT** corrects for this by computing principal components of the genotype matrix, which capture ancestry, and including them as covariates (Price et al. 2006). **Linear mixed models** go further by modelling the genetic relatedness between all pairs of individuals as a random effect, which also handles cryptic relatedness, and scale to biobanks (Loh et al. 2015).

Manhattan and QQ plots. The results are shown as a **Manhattan plot**, with $-\log_{10} P$ for each variant against its position, where associated loci appear as towers of variants in LD with each other. A **quantile–quantile (QQ) plot** compares the observed P-values with the uniform distribution expected under the null hypothesis. A curve that leaves the diagonal only at the tail indicates true associations; a curve lifted from the start indicates inflation, measured as the **genomic inflation factor** λ_{GC}, the median observed χ^2 statistic divided by its expected value of 0.455 (Devlin and Roeder 1999). For highly polygenic traits in large samples, true signal also inflates λ_{GC}, and LD score regression separates polygenicity from confounding (Bulik-Sullivan et al. 2015).

5.8.5 From association to causal variants and genes

A GWAS locus is a region, not a variant. Because of LD, dozens of variants are almost equally associated, and the strongest is often not the causal one. **Statistical fine-mapping** computes, from the association statistics and the LD between variants, the posterior probability that each variant is causal, and reports a **credible set**, the smallest set of variants that contains the causal one with, say, 95% probability (Schaid et al. 2018). Methods such as FINEMAP and SuSiE allow several causal variants per locus. Fine-mapping depends on accurate LD, ideally from the study itself, and is helped by combining populations with different LD patterns, and by functional annotation of the variants.

About 90% of GWAS variants lie outside protein-coding sequence, presumably in regulatory elements (Section 6.9), so the target gene is not obvious and is often not the nearest one. **Colocalisation** tests whether the association with a disease and with the expression of a gene (an eQTL, Section 7.5) are likely to be due to the same causal variant rather than to two distinct variants in LD (Giambartolomei et al. 2014). Overlap with open chromatin and histone marks from the methods of Section 5.5 adds further evidence.

5.8.6 Polygenic scores

A **polygenic score** (PGS or PRS) sums the effects of many variants, $\text{PGS} = \sum_i \hat{\beta}_i x_i$, with x_i the number of alleles and $\hat{\beta}_i$ the effect sizes from a GWAS, after pruning or shrinking the effects to account for LD (Choi et al. 2020). Polygenic scores can identify individuals with a risk comparable to that of carriers

of Mendelian variants, but they predict poorly for individuals, and their accuracy depends strongly on ancestry. Most GWAS participants are of European ancestry, and because LD patterns and allele frequencies differ between populations, scores derived in Europeans are several times less accurate in people of African ancestry (Martin et al. 2019). Using them in the clinic without larger, more diverse studies could widen existing health disparities.

5.8.7 Finding Mendelian disease genes by sequencing

Exome sequencing changed the search for Mendelian disease genes. In 2010, the gene for Miller syndrome was found by sequencing the exomes of only four affected individuals (Ng et al. 2009), and exome and genome sequencing have since become first-line tests for rare disease (Bamshad et al. 2011). Each exome contains about 20,000 variants, and each genome four to five million, so the problem is to filter them down to a few candidates:

- **Annotation.** Each variant is annotated with its consequence (missense, stop-gain, frameshift, splice site, synonymous) in every overlapping transcript, by tools such as VEP (McLaren et al. 2016) and SnpEff (Cingolani et al. 2012). A variant can be a missense change in one transcript and intronic in another, so the transcript set (for example MANE Select) must be stated.
- **Population frequency.** A variant that causes a rare, severe disease cannot be common in the general population. Variants above a frequency cut-off in gnomAD (Section 2.9), adapted to the prevalence and inheritance of the disease, are removed.
- **Inheritance model.** For a recessive disease, the candidate gene must carry two damaging alleles, either homozygous or compound heterozygous, which requires phasing (Section 5.4). For a dominant disease one allele suffices.
- **Trios and de novo mutations.** Each person carries roughly 70 new single-nucleotide variants not present in either parent, on average about one in the exome. Counting also indels, structural variants and changes in tandem repeats, which only long-read assemblies detect reliably, a four-generation pedigree gave about 100–200 de novo mutations per transmission, most of them inherited from the father (Porubsky et al. 2025). Sequencing patient and parents (a *trio*) finds these **de novo** variants directly, and they explain a large fraction of severe developmental disorders and intellectual disability (Veltman and Brunner 2012). Apparent de novo calls are, however, enriched for errors, such as low coverage in a parent.
- **Homozygosity mapping.** In consanguineous families, a recessive disease allele lies in a long run of homozygosity inherited from a common ancestor, which narrows the search to these regions (Lander and Botstein 1987).
- **Phenotype-driven prioritisation.** The patient's clinical features are coded with the Human Phenotype Ontology (HPO) (Köhler et al. 2020), and tools such as Exomiser (Smedley et al. 2015) rank the remaining candidates by combining variant scores with the similarity between the patient's phenotype and those of known diseases and model organisms.

Even so, exome sequencing gives a diagnosis in only about a quarter to two-fifths of patients, depending on the cohort; in critically ill neonates and infants, a meta-analysis found a yield of about 40% for both exome and genome sequencing (Yang et al. 2026). The remainder may have variants in non-coding regions, structural variants, repeat expansions, or variants in genes not yet linked to disease, which is why genome sequencing, long reads and RNA sequencing are increasingly added, and why unsolved cases are reanalysed as knowledge grows.

5.8.8 Interpreting variants in the clinic

Finding a candidate is not the same as proving that it causes the disease. Clinical laboratories classify variants with the ACMG/AMP guidelines into five classes, from *benign* to *pathogenic*, by combining

weighted criteria: population frequency, the predicted consequence (null variants in genes where loss of function causes disease), segregation in the family, de novo occurrence, functional assays and computational predictions (Richards et al. 2015). Many variants end up as *variants of uncertain significance*. The databases used in this process are described in Section 2.9, and the computational prediction of variant effects, which counts only as supporting evidence, in Section 10.6.

See also: further explanation in Section D.5.7; related Wikipedia articles in Section E.5.7.

> **Sources**
>
> Written for this book from the primary literature cited in the text.

5.9 Key concepts

A checklist for revision.

- **Sequencing technologies:** Sanger chain termination and its automation; Phred quality scores; pyrosequencing and 454; Illumina sequencing by synthesis; Ion Torrent; PacBio (CLR and HiFi); Oxford Nanopore (simplex and duplex); read lengths, error profiles and cost, and why quoted accuracies date quickly; detection of base modifications; base qualities as calibrated model estimates; random vs. systematic (context-dependent) errors and why consensus removes only the former; long-read transcriptomics (Iso-Seq, direct RNA sequencing)
- **Shotgun sequencing:** clone-by-clone vs. whole-genome shotgun; coverage and the Lander–Waterman model ($c = NL/G$, uncovered fraction e^{-c}); coverage bias; paired-end and mate-pair reads; read length vs. insert size
- **Assembly paradigms:** shortest common superstring and greedy assembly; overlap–layout–consensus; string graphs and transitive reduction; de Bruijn graphs, Eulerian paths, tips and bubbles, the choice of k; long-read assemblers (Canu, Flye, hifiasm); polishing (Racon) and how it can homogenise repeats or erase heterozygous sites; why "de Bruijn for short, overlap for long reads" is a simplification
- **Repeats** as the central problem of assembly; contigs and scaffolds; scaffolding with read pairs, long reads, optical maps and Hi-C
- **Assembly quality:** contiguity, completeness and correctness as separate axes; N50 and NG50; BUSCO; k-mer based evaluation and QV; reference-based evaluation (QUAST) and why differences from a reference are not always errors; Assemblathon; haplotigs, primary/alternate and haplotype-resolved assemblies; chromosome-scale and telomere-to-telomere assemblies; pangenome references
- **Read mapping:** seed and extend; hash-based mappers vs. Burrows–Wheeler/FM-index (Bowtie, BWA); minimizers and chaining (minimap2); mapping quality; SAM/BAM/CRAM and the SAM record (FLAG, CIGAR, MAPQ); 1- vs. 0-based coordinates; quality control (base quality, adapters, duplicates, contamination, coverage, mapping statistics); reference bias, graph genomes and the human pangenome reference, and the biases of graph references
- **Variant calling:** SNPs, indels and structural variants; the VCF record (REF, ALT, QUAL, FILTER, INFO, FORMAT, GT, AD, PL) and indel normalisation; read evidence vs. genotype likelihood vs. genotype call, and why a difference from the reference is not always a variant; genotype likelihoods and priors; joint calling; GATK and haplotype-based calling; DeepVariant; Genome in a Bottle benchmarks; SV and CNV signals (depth, discordant pairs, split reads, assembly) and what each can detect; long-read SV calling; phasing (read-backed, trio, statistical); minimum error correction (WhatsHap); phase blocks and switch errors
- **Targeted sequencing, copy number and cancer:** hybrid capture and exome sequencing; uneven coverage and off-target reads; CNVs from read depth (binning, GC correction, panel

of normals, segmentation) and from SNP arrays (log R ratio, B-allele frequency, copy-neutral LOH); tumour–normal somatic calling (Mutect); variant allele fraction, purity and ploidy; panel of normals and artefacts; clonal vs. subclonal mutations; drivers vs. passengers, background mutation rate and dN/dS; mutational signatures (96 trinucleotide contexts, NMF, COSMIC signatures)
- **Regulatory and epigenomic sequencing:** ChIP-seq and the input control; peak calling (MACS), narrow vs. broad peaks; IDR and FRiP; motif analysis of peaks; DNase-seq and ATAC-seq, footprints; bisulfite sequencing, bisulfite-aware alignment (Bismark), methylation level and its pitfalls; long-read methylation calling; CLIP-seq
- **Gene prediction:** signals and content; ORFs; why eukaryotic gene prediction is more than ORFs plus splice sites (UTRs, isoforms, non-coding genes); Markov models of coding sequence (GeneMark, Glimmer, Prodigal); coding log-likelihood scores; self-training on long ORFs and its bias against short genes; generalised HMMs (GENSCAN, AUGUSTUS); evaluation at nucleotide, exon and gene level; repeat masking; evidence-based annotation (RNA-seq, protein alignments, MAKER, BRAKER); annotation edit distance; circular evidence and why absence from an annotation is not absence from the genome; deep-learning gene finders
- **Metagenomics:** history (Pace, Handelsman, Sargasso Sea, MetaHIT, HMP, Tara Oceans); 16S amplicon sequencing, OTUs vs. ASVs, diversity; shotgun metagenomics; taxonomic profiling (Kraken, MetaPhlAn) and functional profiling; metagenome assembly; profiling vs. assembly and binning as different questions; binning by composition and coverage; MAG quality (completeness, contamination, MIMAG); MAGs vs. isolate genomes; CAMI
- **Human genetics:** linkage disequilibrium (D, D', r^2), recombination hotspots, haplotype blocks and tag SNPs; Mendelian vs. complex disease, penetrance, allele frequency vs. effect size; heritability and missing heritability; linkage analysis and LOD scores; GWAS design, arrays and imputation; the 5×10^{-8} threshold; population stratification, PCA and mixed models; Manhattan and QQ plots, genomic inflation; replication and winner's curse; fine-mapping and credible sets; colocalisation with eQTLs; polygenic scores and their portability across ancestries; exome/genome sequencing for rare disease: annotation per transcript (VEP), gnomAD frequency filtering, inheritance models, trios and de novo mutations, homozygosity mapping, HPO and Exomiser; ACMG/AMP classification

5.10 Recommended reading

These papers are recommended for reading in full:

- Goodwin et al. (2016) — ten years of next-generation sequencing; how the technologies work and what their errors are.
- Compeau et al. (2011) — how to apply de Bruijn graphs to genome assembly; the key algorithmic idea, very clearly explained.
- Langmead et al. (2009) — Bowtie; read mapping with the Burrows–Wheeler transform.
- Nielsen et al. (2011) — genotype and SNP calling from sequencing data; the probabilistic framework.
- Yandell and Ence (2012) — a beginner's guide to eukaryotic genome annotation.
- Quince et al. (2017) — shotgun metagenomics from sampling to analysis.

Titles, abstracts, links and the full texts are in Section A.5; a longer list of further papers is in Section B.5.

5.11 Background reading (open access)

- **DNA sequencing (Wikipedia)** — DNA sequencing is the process of determining the nucleic acid sequence – the order of nucleotides in DNA. It includes any method or technology that is

used to determine the order of the four bases: adenine, thymine, cytosine, and guanine.
https://en.wikipedia.org/wiki/DNA_sequencing

- **Massive parallel sequencing (Wikipedia)** — Massively parallel sequencing (MPS) is any of several high-throughput approaches to DNA sequencing using the concept of massively parallel processing; it is also called next-generation sequencing (NGS) or second-generation sequencing.
https://en.wikipedia.org/wiki/Massive_parallel_sequencing
- **Nanopore sequencing (Wikipedia)** — Nanopore sequencing is a third generation approach used in the sequencing of biopolymers — specifically, polynucleotides in the form of DNA or RNA.
https://en.wikipedia.org/wiki/Nanopore_sequencing
- **Sequence assembly (Wikipedia)** — In bioinformatics, sequence assembly refers to aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence.
https://en.wikipedia.org/wiki/Sequence_assembly
- **De Bruijn graph (Wikipedia)** — In graph theory, an n-dimensional De Bruijn graph of m symbols is a directed graph representing overlaps between sequences of symbols. It has mn vertices, consisting of all possible length-n sequences of the given symbols; the same symbol may appear multiple times in a sequence.
https://en.wikipedia.org/wiki/De_Bruijn_graph
- **N50, L50, and related statistics (Wikipedia)** — In computational biology, N50 and L50 are statistics of a set of contig or scaffold lengths. The N50 is similar to a mean or median of lengths, but has greater weight given to the longer contigs. It is used widely in genome assembly, especially in reference to contig lengths within a draft assembly.
https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics
- **Burrows–Wheeler transform (Wikipedia)** — The Burrows–Wheeler transform (BWT) rearranges a character string into runs of similar characters, in a manner that can be reversed to recover the original string.
https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform
- **A tutorial on Burrows–Wheeler indexing methods** — A step-by-step blog tutorial (The Grand Locus) on the Burrows–Wheeler transform, the FM-index and backward search, the data structures behind read mappers such as BWA and Bowtie.
http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods
- **SNV calling from NGS data (Wikipedia)** — SNV calling from NGS data is any of a range of methods for identifying the existence of single nucleotide variants (SNVs) from the results of next generation sequencing (NGS) experiments.
https://en.wikipedia.org/wiki/SNV_calling_from_NGS_data
- **Structural variation (Wikipedia)** — Genomic structural variation is the variation in structure of an organism's chromosome, such as deletions, duplications, copy-number variants, insertions, inversions and translocations.
https://en.wikipedia.org/wiki/Structural_variation
- **Gene prediction (Wikipedia)** — In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes. This includes protein-coding genes as well as RNA genes, but may also include prediction of other functional elements such as regulatory regions.
https://en.wikipedia.org/wiki/Gene_prediction
- **Metagenomics (Wikipedia)** — Metagenomics is the study of all genetic material from all organisms in a particular environment, providing insights into their composition, diversity, and functional potential.
https://en.wikipedia.org/wiki/Metagenomics

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

5.12 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

5.13 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Sequencing, Mapping, Assembly, and Variation" and "Genome Architecture, Representation, and Annotation" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

5.13.1 Design sequencing for a repeat-rich diploid genome

You have a highly heterozygous, repeat-rich diploid organism with a genome of 1.2 Gb and no close reference. How would you design the genome project?

Extensions, solution outline and common errors

Extensions:

1. Why are accurate short reads alone insufficient?
2. How do read length and molecule length differ?
3. Would you use Hi-C, parental data or both?
4. What do you mean by "the genome": which output do you deliver?
5. How would you know that both haplotypes were recovered?

Solution outline:

- Define the deliverable first: a collapsed primary assembly, a primary assembly with alternate haplotigs, or two haplotype-resolved assemblies. Estimate genome size, ploidy, heterozygosity and repeat content, for example from a k-mer spectrum.
- Extract high-molecular-weight DNA; long-read sequencing cannot recover long-range information from sheared molecules.
- Use accurate long reads (HiFi) to span repeats and link nearby heterozygous sites. Add parental reads for trio binning if the parents are available, or Hi-C for chromosome-scale phasing and scaffolding; ultra-long nanopore reads help with the longest repeats.
- Assemble with a haplotype-aware method such as hifiasm. Keep bubbles in the graph until the evidence shows whether they are alleles, repeat copies or errors.
- Polish each haplotype separately, so that alleles are not mixed, and be careful with short reads in paralogous regions.
- Evaluate contiguity (N50, NG50), consensus quality (k-mer QV), assembly size relative to the expected genome size, gene completeness (BUSCO), structural correctness (Hi-C contact maps, read support at joins), phasing (haplotype-specific k-mers, switch errors) and contamination.
- Report unresolved regions and gaps rather than forcing a single path.

Common errors:

- Saying “30× coverage” without stating whether it is per haplotype, raw or mapped, and what the read-length distribution is.
- Treating scaffold N50 as proof that the chromosomes are correct.
- Treating all bubbles as sequencing errors.
- Calling a primary assembly “haploid” when it is a mosaic of both homologues.

Connections: Section 5.3; repeats and genome organisation in Section 6.7 and Section 6.4; phasing in Section 5.4.

5.13.2 Explain read mapping

How does a read mapper find the location of a read, and what does mapping quality mean? Avoid simply saying that “the read matches the reference”.

Extensions, solution outline and common errors

Extensions:

1. Compare an FM-index mapper with a minimizer-based mapper.
2. Why is mapping quality not the same as base quality?
3. What is a supplementary alignment?
4. Can a unique alignment still be wrong?
5. What happens if the sample contains sequence that is absent from the reference?

Solution outline:

- Candidate generation: backward search in an FM-index, or seed look-up in a hash table, narrows the reference down to a few positions that are worth an expensive alignment.
- In minimizer mapping (minimap2), representative k-mers chosen from windows become anchors; compatible anchors are chained by order and spacing, and the chain is then aligned base by base.
- The alignment score measures fit under the substitution and gap costs. Mapping quality is a Phred-scaled estimate of the probability that the reported position is wrong, and its computation is specific to each mapper.
- Base quality concerns the decoding of a base; mapping quality concerns the location of the whole read. A read of perfect base quality can map ambiguously, and a unique mapping can contain poor bases.
- Secondary alignments are alternative placements of the same read; supplementary alignments are the parts of a chimeric or rearranged read that map to different places.
- A unique alignment can still be wrong if the heuristic missed alternatives or if the true origin is missing from the reference.
- Diagnose with pair consistency, soft clipping, alternative references, graph mapping and local assembly.

Common errors:

- Interpreting MAPQ 60 as a universal empirical error rate of 10^{-6}.
- Removing all multi-mapping reads and then concluding that a difficult locus has no coverage.
- Assuming that an indexed search evaluates every gapped alignment.
- Taking a high alignment score as proof of biological origin.

Connections: Section 5.4; exact matching and heuristic search in Section 1.4; dynamic programming in Section 1.3.

5.13.3 Compare de Bruijn and string graphs

Explain genome assembly with de Bruijn graphs and with overlap and string graphs, including how each fails.

Extensions, solution outline and common errors

Extensions:

1. What does changing k do in a de Bruijn graph?
2. Why are transitive overlaps removed in a string graph?
3. What are tips and bubbles?
4. When is a repeat resolvable?
5. Why might the graph be a more truthful result than the contigs?

Solution outline:

- In a de Bruijn graph, reads are broken into k-mers; in the usual form, $(k-1)$-mers are nodes and observed k-mers are directed edges. Non-branching paths are compacted into unitigs.
- A small k keeps the graph connected where coverage is low but collapses more repeats; a large k resolves more repeats but breaks at errors and coverage gaps. Assemblers such as SPAdes combine several values of k.
- In overlap–layout–consensus, reads are nodes connected by overlaps. A string graph removes contained reads and transitive edges, which carry no extra information.
- Long reads reduce the number of nodes and span repeats, but overlap detection must tolerate errors.
- Tips may be errors or real sequence with low coverage; bubbles may be errors, heterozygous alleles, strains or paralogues. Removing them is an inference, not housekeeping.
- A repeat is resolved when reads connect each copy to its own unique flanks. Average read length alone does not show this.
- If several paths are equally supported, the graph represents the uncertainty correctly; choosing one contig invents information.

Common errors:

- Saying that de Bruijn nodes are always k-mers without stating the convention used.
- Equating high coverage with repeat resolution.
- Assuming that every long-read assembler literally performs classical overlap–layout–consensus.
- Removing all low-coverage branches in a metagenome, where they may be rare strains.

Connections: Section 5.3; metagenome assembly in Section 5.7; repeats in Section 6.7.

5.13.4 Diagnose an impressive but suspicious assembly

Assembly A has an N50 of 80 Mb and assembly B an N50 of 20 Mb. Is A better?

Extensions, solution outline and common errors

Extensions:

1. Define N50 and NG50 precisely.
2. How can N50 be inflated?
3. What would Merqury and BUSCO tell you?
4. How do you detect a collapsed or duplicated haplotype?
5. How would you validate a chromosome-scale scaffold?

Solution outline:

- N50 is the length such that contigs of that length or longer contain half of the assembly; NG50 uses half of an independent estimate of the genome size instead.
- N50 measures only contiguity. Chimeric joins, collapsed repeats, missing small chromosomes or duplicated haplotigs can all increase it.
- Compare the assembly size with independent genome-size estimates and inspect the *k*-mer spectrum. Merqury compares the *k*-mers of the assembly with those of accurate reads to estimate consensus quality (QV), completeness and, with parental data, phasing.
- BUSCO measures the recovery and duplication of conserved single-copy genes, but says nothing about intergenic sequence or structural correctness.
- Test the structure with read support across joins, Hi-C contact maps, optical or genetic maps, and comparison with a reference where a suitable one exists.
- For haplotypes, examine haplotype-specific *k*-mers, read depth, allele balance and switch errors: a collapsed region has double depth, a duplicated one half depth.
- Conclude only after comparing all these dimensions in relation to the intended use.

Common errors:

- Defining N50 relative to the known genome size rather than the assembly size.
- Calling every difference from a related reference a misassembly.
- Interpreting duplicated BUSCO genes as either always an error or always biological.
- Evaluating a polished consensus only with the reads used for polishing.

Connections: Section 5.3; genome duplication in Section 3.4 and Section 6.4; circular evaluation in 4.

5.13.5 Call and validate small variants

How is a heterozygous SNP inferred from sequencing reads?

Extensions, solution outline and common errors

Extensions:

1. Where do base quality and mapping quality enter?
2. What does local haplotype assembly add?
3. Why might the allele balance differ from 0.5?
4. How does a learned caller such as DeepVariant change the inference?
5. What is a fair benchmark?

Solution outline:

- Define candidate alleles, compute genotype likelihoods $P(D \mid G)$ from the reads under an error model, and combine them with genotype priors to obtain posterior probabilities $P(G \mid D) \propto P(D \mid G)\,P(G)$.
- Base quality gives the probability of a wrong base; mapping quality the probability that the read comes from elsewhere. Strand, position in the read, duplicates and sequence context also matter.
- Local haplotype assembly (as in GATK HaplotypeCaller) represents nearby substitutions and indels jointly and aligns reads to candidate haplotypes, rather than scoring each pileup column on its own.
- A learned caller estimates the decision function from labelled truth data; its performance depends on how well the technology, coverage and ploidy match its training data.
- Allele balance departs from 0.5 because of reference bias, copy number, tumour purity, mosaicism, sampling, amplification or contamination.
- Benchmark only within the confident regions of a truth set such as Genome in a Bottle, and report precision and recall separately by variant type and genomic context.
- Validate important calls with an assay that does not share the same failure modes.

Common errors:

- Taking depth alone as evidence for a heterozygote.
- Treating genotype quality as clinical pathogenicity.
- Counting calls outside the truth set as false positives, although the truth there is unknown.
- Tuning filters on the same benchmark used for the final score.

Connections: Section 5.4; Bayesian reasoning and evaluation in 4; interpretation of variants in Section 10.6.

5.13.6 Detect a structural variant and phase it

A patient may carry a 12 kb insertion in a segmental duplication. How would you detect, reconstruct, phase and validate it?

Extensions, solution outline and common errors

Extensions:

1. What can paired short reads contribute?
2. Why are split reads helpful but sometimes insufficient?
3. How would reference bias appear?
4. How do long reads enable phasing?
5. Which validation assay would you choose?

Solution outline:

- State why the locus is difficult: a 12 kb novel insertion is longer than a short-read fragment, and the segmental duplication makes the anchors ambiguous.
- Short reads give discordant pairs, a change in depth and split reads at the breakpoints, but cannot reconstruct the insertion and may not place it without unique flanks.
- Obtain long reads that span unique sequence on both sides. Keep supplementary alignments and assemble the locus locally rather than trusting a single alignment.
- Compare linear and graph references; check whether reads that support the insertion get low mapping quality or are lost.

- Phase the insertion by linking it to flanking heterozygous sites on the same reads, and report the support for the phase block.
- Reconstruct the inserted sequence, its orientation and copy number, and the breakpoints.
- Validate with long-range PCR or targeted sequencing if possible, or optical mapping for large structures.

Common errors:

- Reporting a symbolic VCF allele as a complete reconstruction.
- Using only total read depth in a duplicated locus.
- Claiming an exact length from reads that end inside the insertion.
- Using a second aligner on the same short reads as the only validation.

Connections: Section 5.4; segmental duplications in Section 6.7; gene duplication in Section 3.4.

5.13.7 Analyse a low-biomass metagenome

Design a shotgun metagenomics study of a clinical sample with little microbial DNA, and explain how you would defend the claim that a rare pathogen is present.

Extensions, solution outline and common errors

Extensions:

1. Why are extraction blanks essential?
2. Compare read classification with assembly-first analysis.
3. What does a lowest-common-ancestor label from Kraken mean?
4. Why does a single mean coverage make little sense?
5. When is a metagenome-assembled genome credible?

Solution outline:

- Randomise cases and controls across extraction and library batches; include extraction and library blanks, positive controls and a mock community.
- Measure the fraction of host DNA. With little microbial DNA, contaminating DNA from reagents and cross-talk between samples matter much more.
- Read classification depends on the database and on k-mers shared between taxa; record the database version and check for reads in regions unique to the organism, not only read counts.
- Assembly can recover novel sequence and genomic context, but uneven abundance, strains, repeats and mobile elements fragment the graph or cause misjoins. Abundance varies over orders of magnitude between organisms, so there is no single coverage.
- A credible bin has coherent composition and abundance across samples, high completeness and low contamination estimated from marker genes, and sufficient read support; it is still a reconstruction.
- For a rare pathogen, require a much weaker signal in the controls, reads from several organism-specific regions, and confirmation with a targeted assay.
- Absence of classified reads is not proof of absence.

Common errors:

- Removing every organism found in any blank without considering its abundance.

- Treating database labels as direct observations of living organisms.
- Removing host reads without documenting the host reference used.
- Reporting species-level abundance when only conserved genus-level sequence is observed.

Connections: Section 5.7; sequence databases in Section 2.3; horizontal gene transfer in Section 3.7.

5.13.8 Decide whether to use a pangenome graph

What problem does a pangenome graph solve, and how would you show that it improves variant calling?

Extensions, solution outline and common errors

Extensions:

1. What are the nodes, edges and paths of the graph?
2. How does a graph reduce reference bias?
3. Can a graph introduce new biases?
4. How are coordinates reported?
5. How would you avoid testing on genomes that were used to build the graph?

Solution outline:

- A single linear reference favours one sequence and lacks sequence specific to other populations. A pangenome graph stores sequences in nodes and alternative adjacencies in edges; paths record the observed haplotypes and the reference.
- Reads with non-reference alleles can align to a matching path instead of paying mismatch and gap penalties or remaining unmapped.
- The graph still depends on which samples and variants were included. Under-represented ancestries and rare sequence remain at a disadvantage, and many similar paths can make mapping more ambiguous.
- Record the graph release, the haplotypes it contains, how coordinates are projected to a linear reference, and the versions of the mapper and caller.
- Compare with a linear workflow on the same reads, and report precision and recall separately by ancestry, region, variant class and repeat context.
- Hold out samples that were not used to build the graph to measure generalisation.
- Conclude that a graph is a broader model of the reference, not an unbiased truth.

Common errors:

- Describing the pangenome as a list of variants without explaining paths and sequences.
- Testing only on individuals that are already represented in the graph.
- Claiming that graph mapping removes all reference bias.
- Ignoring the consequences for coordinates and reproducibility.

Connections: Section 5.4; genome variation within species in Section 6.6; data leakage in 4.

5.13.9 A workflow for prokaryotic genome annotation

A high-contiguity bacterial assembly contains one main chromosome and two circular contigs that may be plasmids. Outline an annotation and validation workflow, and state the evidence needed to trust

each major kind of prediction.

Extensions, solution outline and common errors

Extensions:

1. Why is a long open reading frame not necessarily a gene?
2. What is the role of the coding and background models in the score of a gene?
3. How do short genes, horizontally acquired genes and a wrong genetic code affect the prediction?
4. How should operons and pseudogenes be inferred?
5. How would you distinguish plasmids from circular chromosomal contigs or mobile elements?

Solution outline:

- Fix the assembly version, sample information and genetic code before predicting genes. Check contamination, depth, topology and read support for each replicon.
- For a candidate sequence x, the log-likelihood ratio $\log[P(x \mid C)/P(x \mid N)]$ compares a coding model C with a non-coding model N. Start codons, ribosome-binding sites, length, strand and overlaps add further terms, and dynamic programming selects a compatible set of genes.
- Add protein-family profiles, models for structural RNAs, annotation of mobile elements, and explicit rules for partial genes, pseudogenes, selenoproteins and programmed frameshifts.
- Infer transcription units from strand, spacing, promoters and terminators, conserved gene order and transcript data, and report them as predictions with evidence rather than as certain operons.
- Assess plasmids by replication and partition genes, mobility genes, copy number, topology, read linkage and comparison with known plasmids; circularity alone is not enough.
- Validate representative genes with curated homologues, ribosome profiling or proteomics, transcript boundaries, conserved gene order and manual inspection.

Common errors:

- Reporting every open reading frame from start to stop codon as a gene.
- Assuming that all bacteria use the same ribosome-binding signal.
- Labelling a region with unusual composition as contamination without phylogenetic or read evidence.
- Treating adjacency as proof of co-transcription.
- Naming a protein from a single best hit without checking alignment coverage and domain architecture.

Connections: Section 5.6; profile HMMs in Section 1.6; horizontal gene transfer in Section 3.7; proteomic evidence in Chapter 8; functional annotation in Section 10.4.

5.13.10 Annotating a non-model eukaryotic genome

Design a workflow for the structural annotation of a repeat-rich eukaryotic genome, with stranded short-read RNA-seq, a smaller set of long-read transcripts and proteins from related species as evidence.

Extensions, solution outline and common errors

Extensions:

1. How should repeat annotation precede and inform gene prediction?
2. Explain the role and assumptions of a generalised hidden Markov model.
3. Compare the information from short and long transcript reads.
4. What evidence would you require to keep an alternative isoform?
5. How would you validate the annotation without a complete reference annotation?

Solution outline:

- Start with quality control of the assembly, removal of contamination and organelle sequence, discovery of repeats for the lineage, and a documented choice of soft or hard masking.
- A generalised HMM represents the genome as a path through intergenic, untranslated, exon and intron states, with splice-site and coding signals, length distributions for each segment and reading-frame constraints; the Viterbi algorithm finds the most probable path under the model.
- Short reads measure expression and support individual splice junctions; long reads connect distant exons and show transcript ends, but can be truncated and have platform-specific errors. Aligned proteins from related species stabilise the coding structure.
- Train on complete, high-confidence genes and keep independent genes or evidence for evaluation. Keep the evidence tracks when combining ab initio, transcript and protein models.
- Keep isoforms only when distinct splice junctions, transcript ends, protein products or comparative evidence support them; flag partial, conflicting and unsupported models and those overlapping repeats.
- Report performance at nucleotide, exon, transcript and gene level on curated subsets, together with agreement between evidence types, BUSCO completeness and contamination.

Common errors:

- Hard-masking every repeat and so removing real exons.
- Treating the Viterbi path as biological truth.
- Turning every fragment of a transcript assembly into a gene model.
- Training and evaluating on the same evidence.
- Reporting only a BUSCO score.

Connections: Section 5.6; repeats in Section 6.7; RNA-seq in Section 7.3; data leakage in 4.

5.13.11 Auditing claims of annotation completeness

A genome release reports 98.5% BUSCO completeness, and RNA-seq support for 95% of the predicted genes. Do these figures justify the claim that the annotation is complete? What would a more informative validation report contain?

Extensions, solution outline and common errors

Extensions:

1. What does a benchmark of conserved single-copy orthologues measure?

2. How can RNA-seq support an incorrect gene model?
3. At which levels should precision and recall be measured?
4. How would you test a claim that a gene is absent?
5. How should uncertainty be represented in the released annotation?

Solution outline:

- BUSCO measures the recovery and apparent copy number of a set of genes that are single-copy orthologues in the lineage (taken from OrthoDB). It says nothing about isoform structure, lineage-specific genes, repeats, regulatory elements, chromosome completeness or contamination.
- RNA-seq reads can cover part of a locus while the strand, exon boundaries, merged genes or isoform structure are wrong. Highly expressed genes dominate, and evidence used to build the models is not independent validation.
- Report precision and recall at the level of nucleotides, exact exons, exact transcripts and genes, on curated or held-out data, together with protein evidence, partial genes, overlap with repeats and manual checks.
- To test an apparent absence, search the translated genome sequence and the raw reads with sensitive profiles, examine the syntenic region (Section 6.5), check for assembly gaps, and distinguish deletion, pseudogenisation, divergence and annotation failure.
- Release evidence tracks, confidence levels, stable identifiers, assembly and annotation versions, software and database versions, and known limitations.
- Completeness has many dimensions; the reported percentages support only specific, conditional claims.

Common errors:

- Treating completeness as a single number.
- Counting any overlapping read as support for a transcript.
- Evaluating with the same data that were used for training.
- Interpreting every duplicated BUSCO gene as a real duplication.
- Removing unsupported predictions instead of representing confidence explicitly.

Connections: Section 5.6; assembly quality in Section 5.3; orthology databases in Section 3.4; benchmarking in Section 1.11.

5.13.12 Interpret a genome-wide association study

A case–control GWAS of 20,000 cases and 30,000 controls, genotyped on an array and imputed, reports eight loci with $P < 5 \times 10^{-8}$, a genomic inflation factor $\lambda_{GC} = 1.15$, and a lead SNP at the strongest locus that lies in an intron of gene X. The authors conclude that gene X causes the disease, and propose a polygenic score for clinical screening. How would you evaluate these claims?

Extensions, solution outline and common errors

Extensions:

1. Why is the significance threshold 5×10^{-8}, and does it depend on the number of imputed variants?
2. Is $\lambda_{GC} = 1.15$ evidence of confounding?
3. How would you go from the lead SNP to a causal variant and a causal gene?

4. What does the QQ plot add to the Manhattan plot?
5. In which populations would the polygenic score work?

Solution outline:

- The threshold corrects for about one million independent common variants in European-ancestry genomes; imputation adds variants but few independent tests. Loci should be replicated in an independent cohort, with effect sizes expected to shrink (winner's curse).
- Inflation can come from population stratification, cryptic relatedness or batch effects between case and control genotyping, but also from true polygenic signal in a large sample. Check that principal components or a mixed model were used, and use LD score regression to separate the two.
- The lead SNP marks a region of LD. Statistical fine-mapping gives a credible set of candidate variants; their overlap with regulatory elements in relevant cell types, and colocalisation with eQTLs, suggests the target gene, which may not be X. Experimental perturbation is needed to show causality.
- A QQ plot shows whether the whole distribution of test statistics is inflated, or only its tail.
- A score derived in one ancestry is less accurate in others because of differences in LD and allele frequencies; individual-level prediction is usually weak even within the training population, so clinical value must be shown in independent, diverse cohorts.

Common errors:

- Equating the nearest gene, or the gene containing the lead SNP, with the causal gene.
- Treating the lead SNP as the causal variant.
- Interpreting every $\lambda_{GC} > 1$ as confounding, or ignoring it.
- Confusing a strong P-value with a large effect size.
- Applying a polygenic score across ancestries without validation.

Connections: Section 5.8; eQTLs in Section 7.5; regulatory elements in Section 6.9; the GWAS Catalog in Section 2.9.

5.13.13 Find the cause of a rare disease in a trio

A child has a severe developmental disorder; the parents are healthy and unrelated. Exome sequencing of the child and both parents gives about 25,000 variants per person. Describe how you would find the causal variant, and how you would decide whether a candidate should be reported.

Extensions, solution outline and common errors

Extensions:

1. Which inheritance models are plausible, and how does each use the parents' data?
2. Which population frequency cut-off would you use?
3. Why might an apparent de novo variant be an artefact?
4. How would the analysis change if the parents were first cousins?
5. What would you do if no candidate is found?

Solution outline:

- Check sample identity and relationships, coverage and call quality. Annotate variants per transcript with VEP or SnpEff, using a stated transcript set.

- With healthy, unrelated parents, consider de novo dominant variants (present in the child, absent from both parents), recessive variants (homozygous, or compound heterozygous with one allele from each parent, which the trio phases) and X-linked variants in a boy.
- Remove variants that are too common in gnomAD for a rare, severe disease (for a de novo dominant model, essentially absent), keep protein-altering and splice variants, and rank them by predicted effect, gene constraint and phenotype match using HPO terms and tools such as Exomiser.
- Check de novo candidates for adequate depth in both parents, mapping quality, and strand and position biases; confirm by an independent method.
- Classify the best candidate with the ACMG/AMP criteria: frequency, consequence, de novo status, segregation, functional data and computational evidence; report pathogenic and likely pathogenic variants in genes consistent with the phenotype, and handle variants of uncertain significance with care.
- With consanguinity, prioritise long runs of homozygosity. If nothing is found, consider genome sequencing for non-coding, structural and repeat-expansion variants, RNA sequencing, and periodic reanalysis.

Common errors:

- Treating absence from gnomAD as proof of pathogenicity.
- Accepting de novo calls without checking parental coverage.
- Ignoring the transcript in which a variant is annotated.
- Using computational predictions as the main evidence.
- Assuming that a negative exome excludes a genetic cause.

Connections: Section 5.8; variant calling and phasing in Section 5.4; variant databases in Section 2.9; variant effect prediction in Section 10.6.

6 Genomes

> **From the exam description**
>
> Fundamental differences in how different genomes are constructed. What is the difference between eukaryotic and prokaryotic genomes?

6.1 Overview

This chapter summarises what thirty years of genome sequencing have taught us about genomes: how large they are, how they are organised, how they vary within species, and what the non-coding majority of a eukaryotic genome does. The methods for obtaining genome sequences are covered in Chapter 5, and the evolutionary processes that shape genomes, such as duplication, horizontal transfer and selection, in Chapter 3. Gene expression, the topic of the next chapter (Chapter 7), builds on the regulatory elements introduced here.

- background: nucleic acids, the central dogma, cells and model organisms (Section 6.2);
- the genome projects, what a gene is and why the human gene count fell, genome size and gene number, and the C-value paradox (Section 6.3);
- the organisation of prokaryotic and eukaryotic genomes, genome duplications, organelle genomes and viral genomes (Section 6.4);
- gene order, synteny, dot plots and genome rearrangements (Section 6.5);
- genomes in populations: pan-genomes, from one reference genome to a pangenome, population-genetic explanations of genome complexity, and what ancient genomes have revealed about human evolution (Section 6.6);
- repeats, transposable elements, segmental duplications, the mechanisms that create structural variants, and the 3D organisation of the genome (Section 6.7);
- pseudogenes, the birth of new genes (including de novo from non-coding DNA), and small proteins missed by classical annotation (Section 6.8);
- non-coding RNAs, regulatory elements and the debate about how much of the genome is functional (Section 6.9).

6.2 Background: genes, genomes and cells

Bioinformatics at DBB is part of a PhD in biochemistry, and examiners expect you to know the molecules and cells that the methods describe. This section summarises the background on nucleic acids and genes, at the level of an undergraduate course in molecular biology. The corresponding background on proteins is given in Section 9.3.

6.2.1 Nucleic acids

DNA is a polymer of four nucleotides, each consisting of a deoxyribose sugar, a phosphate and one of four bases: the purines adenine (A) and guanine (G), and the pyrimidines cytosine (C) and thymine (T). Two antiparallel strands form the **double helix**, held together by **base pairing**, A with T (two hydrogen bonds) and G with C (three), and by stacking between the bases. Because the strands are complementary, each can serve as a template for the other, which is the basis of replication and of almost every experimental method in molecular biology: hybridisation, PCR and sequencing (Section 5.2). RNA differs by having a ribose sugar and uracil (U) instead of thymine. It is usually single-stranded but folds into secondary and tertiary structures through base pairing within the same molecule (Section 1.7), which allows it to act as a catalyst as well as a carrier of information (Section 3.2).

6.2.2 The central dogma and its exceptions

Francis Crick's **central dogma** describes the flow of sequence information in cells: DNA is **replicated**, **transcribed** into RNA, and messenger RNA is **translated** into protein. In eukaryotes, the primary transcript is **processed** before export from the nucleus: a 5' cap is added, introns are removed by **splicing** (Section 6.4), and a poly(A) tail is added at the 3' end. Translation by the ribosome reads the mRNA in triplets, using the **genetic code**, in which 61 codons specify the 20 amino acids and three are stop signals. The code is nearly universal, with small variations, for example in mitochondria. Because most amino acids are encoded by several codons, substitutions can be **synonymous** or **non-synonymous**, the basis of the dN/dS test for selection (Section 3.6).

The dogma has important exceptions and extensions. **Reverse transcription** copies RNA into DNA, in retroviruses and retrotransposons (Section 6.7). Many genes produce functional RNAs that are never translated (Section 6.9). Gene expression is regulated at every step, by transcription factors binding to promoters and enhancers, by chromatin, by RNA processing and stability, by microRNAs and by the regulation of translation and protein degradation, which is why mRNA levels predict protein levels only partly (Section 7.5).

6.2.3 Cells and organisms

All cellular life belongs to three domains, Bacteria, Archaea and Eukarya (Section 4.5). **Prokaryotic** cells (bacteria and archaea) have no nucleus, and transcription and translation are coupled in the cytoplasm. **Eukaryotic** cells have a nucleus, an endomembrane system (endoplasmic reticulum, Golgi apparatus, lysosomes) and organelles: **mitochondria**, and in plants and algae **chloroplasts**, which descend from bacterial endosymbionts (Section 6.4). Proteins are sorted to these compartments by signals in their sequence (3).

Much of what is known comes from a few **model organisms**, chosen because they are easy to grow and manipulate genetically: the bacterium *Escherichia coli*, the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, the zebrafish and the mouse. Their genomes were among the first sequenced (Section 6.3), and knowledge about their genes is transferred to other species by homology (Section 10.4).

6.2.4 Experimental methods

You should also understand, in outline, the experimental methods that produce the data you analyse: PCR and DNA sequencing (Section 5.2), microarrays and RNA-seq (Chapter 7), mass spectrometry

(Chapter 8), X-ray crystallography, NMR and cryo-EM (Section 9.4), and the methods for detecting protein interactions (Section 11.2). Knowing how data are generated is essential for knowing their errors and biases.

See also: further explanation in Section D.6.1; related Wikipedia articles in Section E.6.1.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles DNA, Central dogma of molecular biology, Genetic code and Model organism (CC BY-SA 4.0, retrieved 4 October 2026). For more detail, see Alberts et al. (2002) (free on the NCBI Bookshelf) or Clark et al. (2018) (OpenStax, free).

6.3 From the first genome projects

The first complete genomes, in the 1990s, turned genomics from the study of single genes into the comparison of whole gene repertoires. This section follows the main genome projects and summarises what they showed about how genomes differ in size and gene content. The technologies and algorithms behind them are described in Chapter 5.

6.3.1 The genome projects

The first DNA genome, that of the bacteriophage φX174 (5,386 bases), was sequenced in 1977, and the human mitochondrial genome (16,569 bases) followed in 1981 (Anderson et al. 1981). Sequencing a cellular genome seemed out of reach until the 1990s.

Table 6.1: Some landmark genome sequences. Sizes marked * are the euchromatic sequence only.

Year	Genome	Size	Genes	Notes
1995	*Haemophilus influenzae*	1.8 Mb	~1,700	first free-living organism; whole-genome shotgun (Fleischmann et al. 1995)
1995	*Mycoplasma genitalium*	0.58 Mb	~470	one of the smallest genomes of a cell that can be grown in culture; the "minimal gene set" (Fraser et al. 1995)
1996	*Saccharomyces cerevisiae*	12 Mb	~6,000	first eukaryote; international consortium (Goffeau et al. 1996)
1997	*Escherichia coli* K-12	4.6 Mb	~4,300	the model bacterium (Blattner et al. 1997)
1998	*Caenorhabditis elegans*	97 Mb	~19,000	first animal (The C. elegans Sequencing Consortium 1998)
2000	*Drosophila melanogaster*	120 Mb*	~13,600	whole-genome shotgun of a large genome (Adams et al. 2000; Myers et al. 2000)
2000	*Arabidopsis thaliana*	125 Mb	~25,500	first plant (The Arabidopsis Genome Initiative 2000)
2001	*Homo sapiens* (drafts)	2.9 Gb	~30,000–40,000 predicted	public and Celera drafts (International Human Genome Sequencing Consortium et al. 2001; Venter et al. 2001)
2002	*Mus musculus*	2.5 Gb	~30,000 predicted	first comparison of two mammals (Mouse Genome Sequencing Consortium 2002)
2004	*Homo sapiens* (finished)	2.85 Gb*	20,000–25,000	finished euchromatin (International Human Genome Sequencing Consortium 2004)

Year	Genome	Size	Genes	Notes
2022	*Homo sapiens* (T2T)	3.06 Gb	~19,000–20,000	first complete human genome (Nurk et al. 2022)

The **Human Genome Project** was launched in 1990 as an international public project, with an estimated cost of three billion dollars. It used a *hierarchical* (clone-by-clone) strategy: the genome was first cut into large bacterial artificial chromosome (BAC) clones, which were mapped to the chromosomes and then sequenced individually by shotgun. In 1998 the company Celera Genomics, led by Craig Venter, announced that it would sequence the genome faster with a *whole-genome shotgun* strategy (Weber and Myers 1997), relying on the assembly algorithms developed by Gene Myers (Section 5.3). The resulting race ended with the joint announcement of two draft sequences in 2000 and their publication in 2001 (International Human Genome Sequencing Consortium et al. 2001; Venter et al. 2001). One lasting legacy of the public project is the *Bermuda principles* of 1996, which required sequence data to be released within 24 hours, and which set the norm of open data in genomics (Chapter 2).

What counts as a *finished* genome has changed. The "finished" human sequence of 2004 covered about 99% of the euchromatin, with about one error per 100,000 bases, but it still had 341 gaps and left out the heterochromatin (International Human Genome Sequencing Consortium 2004). The reference assembly GRCh38 (2013) remained incomplete in the same places. About 8% of the genome, including the centromeres, the short arms of the acrocentric chromosomes with the ribosomal RNA gene arrays, and large segmental duplications, consists of long, nearly identical repeats (Section 6.7) that could not be assembled from BAC clones or short reads. Accurate long reads combined with ultra-long nanopore reads made it possible to assemble them, and in 2022 the Telomere-to-Telomere (T2T) consortium published the first gapless sequence of all human autosomes and the X chromosome, CHM13 (Nurk et al. 2022; Altemose et al. 2022); a complete Y chromosome followed in 2023 (Rhie et al. 2023) (Section 5.3). A finished genome now means a gapless sequence of every chromosome from telomere to telomere, ideally with both haplotypes resolved.

The early genomes produced several surprises. About a third of the genes in every new bacterial genome had no known function. Even *E. coli* and yeast, the best-studied organisms, had hundreds of uncharacterised genes. The human genome contained far fewer genes than expected: estimates before the project ranged up to 100,000, but the finished sequence suggested 20,000–25,000 (International Human Genome Sequencing Consortium 2004), and the current count of protein-coding genes is around 19,000–20,000 (as of 2026, depending on the annotation). This is about the same as in the nematode *C. elegans*. The high early estimates rested on the expectation that a more complex organism needs more genes; the genome projects showed that this expectation was wrong (see the G-value paradox below). The mouse genome showed that about 5% of the human genome is conserved between the two species, much more than the 1–2% that codes for protein (Mouse Genome Sequencing Consortium 2002), which gave the first evolutionary estimate of how much of the genome is functional (Section 6.9).

Gene counts come from a genome **annotation**, which is a separate product from the genome **assembly**, with its own versions and its own errors. The assembly is the sequence itself, released in versions such as GRCh37, GRCh38 (with minor patch releases) and T2T-CHM13; genomic coordinates refer to one assembly version and differ between versions. The annotation is the set of gene models placed on an assembly by pipelines that combine *ab initio* prediction, transcript and protein evidence and manual curation (Section 5.6). For the human genome, GENCODE (identical to the Ensembl annotation) and RefSeq annotate the same assembly independently and are updated several times a year. They differ in the number of genes, in the transcripts of each gene and in exon boundaries, which is why the MANE project selects one matching transcript per protein-coding gene that both agree on. Annotations of less-studied genomes are mostly automatic and contain more errors, such as split, fused, missing or spurious gene models, which propagate to the protein databases (Chapter 2). A result should therefore always state both the assembly and the annotation version.

6.3.2 What is a gene?

Counting genes requires a definition of a gene, and the definition has changed several times (Gerstein et al. 2007). For Mendel and the early geneticists, a gene was an abstract unit of inheritance, detected through a phenotype and placed on a linkage map. The *one gene, one enzyme* idea of the 1940s and the discovery of DNA turned it into a molecular object: a stretch of DNA that encodes one polypeptide. This molecular gene was soon complicated by split genes and alternative splicing (Section 6.4), by genes that overlap or lie inside the introns of other genes, by distant regulatory elements (Section 6.9), and by genes whose final product is an RNA rather than a protein.

The ENCODE pilot project of 2007 added another complication. Much of the genome is transcribed, transcripts often extend beyond the annotated genes or join exons of neighbouring genes, and many transcribed regions overlap no known gene. In response, Gerstein and colleagues proposed that a gene is "a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein et al. 2007). The unit is the *product* (a protein or a functional RNA), and the gene is the set of sequences that contribute to related products. In practice, annotations such as GENCODE are built from transcript models, and genes are clusters of transcripts that share exons on the same strand.

The count of human protein-coding genes has fallen as the evidence has improved (Pertea and Salzberg 2010). Before the genome was sequenced, estimates were often around 100,000, and some estimates based on expressed sequence tags were even higher. The drafts of 2001 predicted 30,000–40,000, the finished genome 20,000–25,000, and the count has drifted down to about 19,000–20,000 since. The reduction came from removing models that do not encode proteins: ORFs that are not conserved and have no protein-like features (Clamp et al. 2007), fragments of pseudogenes, and models without support from protein evidence or evolutionary constraint (Ezkurdia, Juan, et al. 2014). At the same time, the counts of non-coding genes and of transcripts per gene have grown.

The reference catalogues still differ (Amaral et al. 2023). GENCODE/Ensembl and RefSeq agree on most protein-coding genes but not on all, and CHESS, built from thousands of RNA-seq experiments (Pertea et al. 2018), gives other numbers again. The disagreements are concentrated in a few hundred to a few thousand loci: genes of uncertain coding status, which may be non-coding RNAs or pseudogenes (Abascal et al. 2018), readthrough transcripts that join two genes, and complex families in segmental duplications. Differences in how isoforms and UTRs are defined are much larger. MANE Select defines one matching transcript per protein-coding gene in Ensembl and RefSeq (Morales et al. 2022), which makes clinical reporting consistent but does not settle which loci are genes. Small protein-coding genes are a separate blind spot, discussed in Section 6.8. Whether a locus is "a gene" is therefore partly a decision of the annotation, and a gene count is only meaningful with the catalogue and release it came from.

6.3.3 Genome size and gene number

The **C-value** is the amount of DNA in a haploid (gametic) nucleus. It varies more than 100,000-fold among cellular organisms:

- **Bacteria and archaea** have genomes from about 0.1 to 16 Mb. The smallest belong to endosymbionts of insects, such as *Carsonella*, with 160 kb and fewer than 200 genes (Nakabachi et al. 2006). They have lost most genes needed for independent life. The smallest free-living organisms, such as the abundant marine bacterium *Pelagibacter*, have streamlined genomes of about 1.3 Mb (Giovannoni et al. 2005). The largest bacterial genomes, in soil bacteria such as myxobacteria and streptomycetes, are 10–16 Mb.

- **Eukaryotes** range from about 2 Mb in some microsporidian parasites to more than 100 Gb. The current record is the fork fern *Tmesipteris oblanceolata*, with about 160 Gb, some 50 times the human genome (Fernández et al. 2024). The lungfish, salamanders and lilies also have very large genomes.

In prokaryotes, genomes are densely packed with genes, about one gene per kilobase, and **gene number is proportional to genome size**. Larger bacterial genomes encode more regulatory and metabolic genes, reflecting more versatile lifestyles. In eukaryotes the relationship breaks down. Yeast has about one gene every 2 kb, while in the human genome protein-coding exons make up only about 1.5% of the sequence.

Genome size and gene number should therefore be kept apart. Among eukaryotes, differences in genome size are mostly differences in repetitive DNA, above all transposable elements (Section 6.7), and in the length of introns and intergenic regions, not in the number of protein-coding genes. **Polyploidy** is a further, independent factor. A recent polyploid, such as bread wheat (about 16 Gb, with three subgenomes), carries several copies of each chromosome set, which multiplies the DNA content and the number of gene copies without adding new kinds of genes (Section 6.4). For polyploids, the C-value, which includes all chromosome sets of a gamete, is therefore distinguished from the *monoploid* genome size, the DNA content of a single chromosome set.

6.3.4 The C-value paradox

Genome size is not correlated with organismal complexity: an onion has about five times more DNA than a human, and closely related species of amphibians or plants can differ several-fold in genome size. This was called the **C-value paradox** by C. A. Thomas in 1971 (Thomas 1971). The paradox was resolved, in part, by the discovery that most eukaryotic DNA does not code for proteins: differences in genome size are mostly differences in non-coding DNA, especially transposable elements (Section 6.7). What remains unexplained is *why* different species carry such different amounts of non-coding DNA, which Gregory called the **C-value enigma** (Gregory 2001). Genome size correlates with cell size, nucleus size and the rate of cell division, and it changes over time through a balance between gains, mainly by transposable elements and duplications, and losses by deletion (Kapusta et al. 2017). Population-genetic arguments for why non-coding DNA accumulates in some lineages are discussed in Section 6.6.

A similar observation applies to genes: the number of protein-coding genes does not track complexity either (sometimes called the *G-value paradox*). Humans have about as many genes as nematodes and fewer than many plants. Organismal complexity is itself hard to define, and the number of protein-coding genes is not a measure of it. The differences between, for example, humans and nematodes are thought to lie more in how genes are used: more elaborate regulation, alternative splicing (Section 6.4), combinations of protein domains (Section 3.5), and non-coding RNAs (Section 6.9).

See also: further explanation in Section D.6.2; related Wikipedia articles in Section E.6.2.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Genome project, Human Genome Project, Genome size, C-value and Bermuda Principles (CC BY-SA 4.0, retrieved 3 October 2026).

6.4 Differences across the tree of life

The exam asks for the fundamental differences between prokaryotic and eukaryotic genomes. This section describes the organisation of each, the traces of whole-genome duplications, the genomes of

mitochondria and chloroplasts, which are remnants of bacteria living inside eukaryotic cells, and the genomes of viruses.

6.4.1 Prokaryotic genome organisation

Most bacteria and archaea have a single **circular chromosome**, often together with smaller circular **plasmids**. There are exceptions: *Vibrio cholerae* has two chromosomes, and *Borrelia* and *Streptomyces* have linear ones. The genome is not enclosed in a nucleus but folded into a compact *nucleoid*. Prokaryotic genomes are organised by replication, transcription and selection for compactness (Rocha 2008):

- **One origin of replication.** Replication starts at a single origin (*oriC*) and proceeds in both directions to the terminus, dividing the chromosome into two halves, *replichores*. Many archaea, in contrast, have several origins.
- **GC skew.** Because the leading and lagging strands are replicated differently, they accumulate different mutations: the leading strand becomes enriched in G over C. The GC skew, $(G-C)/(G+C)$ in a sliding window, therefore changes sign at the origin and the terminus, which is used to locate them in new genome sequences (Lobry 1996).
- **Gene orientation and position.** Genes, especially essential and highly expressed ones, tend to lie on the leading strand, so that the replication fork and RNA polymerase move in the same direction and do not collide head-on. Highly expressed genes, such as those for ribosomal proteins, cluster near the origin, where they are present in more copies in rapidly dividing cells.
- **Operons.** Functionally related genes are often transcribed together as one mRNA from a single promoter. The concept was introduced by François Jacob and Jacques Monod in 1961 with the *lac* operon (Jacob and Monod 1961). About half of bacterial genes are in operons. Operons are rearranged over evolution, but genes that remain neighbours in many genomes are usually functionally linked, which is used to predict function (Chapter 11).
- **Compactness.** About 85–90% of the genome codes for proteins, genes lack introns, intergenic regions are short (around 100 bp), and there are few pseudogenes. Prokaryotic genomes have a *deletional bias*: non-functional DNA is removed quickly.
- **A mobile gene pool.** Plasmids, prophages, transposons and *genomic islands* acquired by horizontal gene transfer make up a variable part of the genome (Section 6.6 and Section 3.7).

Archaea look like bacteria in this respect, with circular chromosomes, operons and compact genomes, but their machinery for replication, transcription and translation resembles that of eukaryotes, and some of them package DNA with histones.

6.4.2 Eukaryotic genome organisation

Eukaryotic genomes are organised very differently:

- **Linear chromosomes in a nucleus.** The genome is divided into several linear chromosomes (23 pairs in humans), packed with histones into chromatin (Section 6.7). Each chromosome is replicated from many origins.
- **Telomeres.** Linear chromosomes cannot be replicated to their very ends (the *end-replication problem*). They end in **telomeres**, tandem repeats of a short motif (TTAGGG in vertebrates) that are added by the enzyme **telomerase**, an RNA-containing reverse transcriptase discovered by Carol Greider and Elizabeth Blackburn (Greider and Blackburn 1985).
- **Centromeres** attach chromosomes to the spindle during cell division. They range from a 125 bp sequence in budding yeast to megabase arrays of satellite repeats in humans (Section 6.7).

- **Introns and splicing.** In 1977 two groups found that adenovirus mRNAs are assembled from separate segments of the genome (Berget et al. 1977; Chow et al. 1977). Most eukaryotic genes are split into **exons** and **introns**, which are removed from the pre-mRNA by the spliceosome. Intron density varies widely: budding yeast has only about 280 introns in total, while a typical human gene has about ten exons spread over tens of kilobases, and some genes, such as dystrophin, span more than 2 Mb. Whether introns were present in the earliest genes (*introns early*) or were inserted later (*introns late*) was long debated. Spliceosomal introns probably derive from self-splicing group II introns that entered with the mitochondrial endosymbiont.
- **Alternative splicing.** Exons can be combined in different ways, so one gene can produce several transcripts. RNA-seq showed that more than 90% of human multi-exon genes are alternatively spliced (Pan et al. 2008; Wang et al. 2008), although how many of these isoforms are functional is debated (Section 3.5).
- **Large intergenic regions and distant regulation.** Genes are separated by long stretches of non-coding DNA, much of it derived from transposable elements. Genes are regulated by promoters and by enhancers that can be far from the gene they control (Section 6.9), and transcription is influenced by chromatin and its 3D organisation (Section 6.7).

Table 6.2: Main differences between prokaryotic and eukaryotic genomes.

Feature	Prokaryotes	Eukaryotes (e.g. human)
Chromosomes	usually one, circular, plus plasmids	several, linear, in a nucleus
Replication origins	one (bacteria); one to a few (archaea)	many per chromosome
Ends	none	telomeres and telomerase
Genome size	0.1–16 Mb	2 Mb – 160 Gb
Coding fraction	85–90%	1–2% in humans (70% in yeast)
Introns	very rare	common; spliceosome
Gene organisation	operons, polycistronic mRNA	one gene per mRNA, alternative splicing
Repeats and transposons	few	often the majority of the genome
Gene content	highly variable within species (pan-genome)	relatively stable within species
Organelle genomes	none	mitochondria, plastids

6.4.3 Genome duplications

Whole-genome duplications (WGDs) have left clear traces in many eukaryotic genomes. Their mechanisms, the fates of duplicated genes and the evidence for the main events are described in Section 3.4. At the level of genome organisation, a WGD is recognised as large blocks of duplicated genes in the same order (**paralogous synteny**), and, when two related species are compared, as a pattern in which each region of one genome corresponds to two regions of the other. The budding yeast genome has such a 2:1 relationship with genomes that did not undergo its WGD about 100 million years ago (Wolfe and Shields 1997; Kellis et al. 2004). Two rounds of WGD early in vertebrate evolution (*2R*) explain why many human gene families have four members where invertebrates have one (Dehal and Boore 2005). Teleost fish had a third round, and flowering plants are particularly rich in polyploidy: the small genome of *Arabidopsis* contains the remnants of several ancient duplications (The Arabidopsis Genome Initiative 2000). Most duplicated genes are lost within a few million years, and genomes return towards their previous size, so ancient WGDs can be detected only through the retained duplicates and the conserved gene order.

Recent polyploids and hybrids are among the hardest genomes to analyse. In an *allopolyploid*, formed by hybridisation between species, as in bread wheat (subgenomes A, B and D) and many other crops, the subgenomes are similar but not identical. In an *autopolyploid*, such as cultivated potato (four copies of one genome), the copies are nearly identical. Assemblers designed for haploid or diploid genomes tend to collapse similar subgenomes or switch between them, and short reads map ambiguously between them, so separating them requires long reads and phasing (Section 5.3). Orthology is harder too. A gene of a diploid relative corresponds to two or more *homoeologues*, the copies inherited from the different parental genomes, so one-to-one orthology breaks down, and gene trees of an allopolyploid reflect hybridisation, a reticulate event, rather than a simple bifurcating species tree (Section 3.4; Section 4.4). Hybrid fungal genomes, such as that of the lager yeast *Saccharomyces pastorianus*, raise the same problems. Most genomics methods and benchmarks were developed on diploid model organisms, and their performance on polyploid and hybrid genomes should not be taken for granted.

6.4.4 Organelle genomes and endosymbiosis

Mitochondria and chloroplasts contain their own small genomes. The **endosymbiotic theory**, proposed for chloroplasts by Konstantin Mereschkowski in 1905 and developed by Lynn Margulis (then Sagan) in 1967 (Sagan 1967), explains them as the remnants of bacteria that were engulfed by an ancestral cell. Gene sequences confirmed the theory: mitochondria descend from alphaproteobacteria and plastids from cyanobacteria (Roger et al. 2017). The acquisition of the mitochondrion was probably part of the origin of eukaryotes themselves (Section 4.5).

- The **human mitochondrial genome** is a circle of 16,569 bp encoding 37 genes: 13 proteins of the respiratory chain, 22 tRNAs and 2 rRNAs (Anderson et al. 1981). It is very compact, has no introns, uses a slightly different genetic code (UGA codes for tryptophan) and is inherited from the mother. Its rapid evolution and lack of recombination made it the first molecular marker for human population history (Section 6.6).
- Mitochondrial genomes vary greatly. The most bacteria-like, in the protist *Reclinomonas americana*, encodes 97 genes (Lang et al. 1997), while plant mitochondrial genomes are large (hundreds of kilobases to megabases) but mostly non-coding.
- **Plastid genomes** are typically 120–160 kb with 100–130 genes.

Organelle genomes encode only a small fraction of the proteins the organelles need. Human mitochondria contain more than 1,000 proteins, almost all encoded in the nucleus, made in the cytoplasm and imported. Most of the genes of the original endosymbionts were lost or moved to the nucleus by **endosymbiotic gene transfer** (Timmis et al. 2004), and the transfer continues: fragments of mitochondrial DNA are still inserted in nuclear genomes (*NUMTs*). Why some genes remain in the organelle at all is still discussed; one explanation is that it allows each organelle to regulate the expression of the genes for its own respiratory or photosynthetic complexes. Endosymbiont genomes in general shrink. The bacterial endosymbionts of insects, which live inside host cells and are transmitted between generations, have some of the smallest genomes known (Nakabachi et al. 2006), a process driven by small population sizes and relaxed selection (Section 6.6).

6.4.5 Viral genomes

Viruses are not cells, but their genomes are the most varied of all. A viral genome can be DNA or RNA, single- or double-stranded, linear or circular, and in one piece or *segmented*, like the eight RNA segments of influenza virus. David Baltimore classified viruses by how they make their mRNA from the genome (Baltimore 1971); the resulting Baltimore classes (double-stranded DNA, single-stranded DNA, double-stranded RNA, positive- and negative-sense single-stranded RNA, and reverse-transcribing viruses with RNA genomes, to which reverse-transcribing DNA viruses were added later) are still used,

although they are not a taxonomy. Viruses share no universal genes, so the formal taxonomy of the International Committee on Taxonomy of Viruses groups them into realms by a few *hallmark* genes, such as those for capsid proteins and replication enzymes (Koonin et al. 2020).

Genome sizes range from under 2 kb in circoviruses to 2.5 Mb in the pandoraviruses of amoebae, larger than some bacterial and even eukaryotic genomes (Philippe et al. 2013). Small viral genomes are extremely compact. **Overlapping genes**, read in different frames of the same sequence, were discovered in the first DNA genome to be sequenced, that of bacteriophage φX174 (Sanger, Air, et al. 1977). Many RNA viruses translate their genome into one long *polyprotein* that is cleaved by viral proteases, and use programmed ribosomal frameshifting or subgenomic mRNAs to express further genes, as coronaviruses do. RNA viruses have mutation rates of about 10^{-6}–10^{-4} per nucleotide per cell infection, far higher than DNA viruses (Sanjuán et al. 2010). This is thought to limit their genome size: coronaviruses, among the largest RNA viruses at about 30 kb, encode a proofreading exonuclease.

These features matter for analysis. Gene prediction based on open reading frames misses overlapping genes, frameshifts and spliced transcripts; segmented genomes can exchange segments (*reassortment*) and RNA viruses recombine, so that different parts of a genome have different trees (Section 4.4); and the high mutation rate means that a sample from one host is a population of variants rather than a single sequence.

See also: further explanation in Section D.6.3; related Wikipedia articles in Section E.6.3.

Sources

Parts of this section are adapted from the Wikipedia articles Prokaryote, Bacterial genome, GC skew, Operon, Eukaryotic chromosome structure, Telomere, Intron, Paleopolyploidy, Symbiogenesis and Mitochondrial DNA (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Genome Architecture, Representation, and Annotation” (CC BY 4.0), used with permission.

6.5 Gene order, synteny and genome rearrangements

Genomes change not only by substitutions, duplications and losses but also by rearrangements that move, invert and join large segments. The order of genes along the chromosomes is therefore a character in its own right. It decays more slowly than sequence in some lineages and much faster in others, and it provides evidence of homology, orthology, genome duplication and function that is independent of sequence similarity.

6.5.1 Synteny and collinearity

The word **synteny** originally meant that two loci lie on the same chromosome. In comparative genomics it is mostly used for conserved correspondence between regions of two genomes, and **collinearity** for the stricter case in which the genes are also in the same order and orientation. Because usage varies, it should be stated whether a claim concerns the same chromosome, a block of conserved gene content, or a collinear block.

From a few dozen genetic markers, Nadeau and Taylor estimated in 1984 that the human and mouse genomes are made up of about 180 conserved segments, broken by rearrangements since the two lineages diverged (Nadeau and Taylor 1984). The mouse genome sequence confirmed the picture: more than 90% of both genomes can be divided into a few hundred corresponding blocks of conserved synteny (Mouse Genome Sequencing Consortium 2002). Rates differ greatly between lineages. Gene order in mammals

and birds is relatively stable, whereas plant mitochondrial genomes rearrange so often that closely related species differ in gene order even though their sequences evolve slowly (Palmer and Herbon 1988). In bacteria, gene order beyond the operon is lost quickly, so that distantly related genomes retain little conserved order except for groups of functionally linked genes (Mushegian and Koonin 1996). Genes that remain neighbours across many bacterial genomes are therefore likely to work together, which is used to predict functional links (Section 10.4).

Conserved synteny is used in several ways:

- **Orthology.** Among several paralogues, the orthologue usually lies in the corresponding region of the other genome. Orthology databases use conserved gene order as independent support (Section 3.4).
- **Missing genes.** If a gene is absent from an otherwise conserved block, a search of the genome sequence in that region often finds a gene that the annotation missed, a pseudogene, or a gap in the assembly.
- **Genome duplications.** A whole-genome duplication leaves pairs of paralogous blocks, and each region of an unduplicated relative corresponds to two regions of the duplicated genome (Section 6.4).
- **Ancestral genomes.** Comparing many genomes makes it possible to reconstruct ancestral chromosomes. The amphioxus genome showed that chordate chromosomes descend from about 17 ancestral linkage groups (Putnam et al. 2008), and conserved synteny has been traced across animals as distant as sponges, cnidarians and bilaterians (Simakov et al. 2022). Patterns of synteny between amphioxus, lampreys and jawed vertebrates have also been used to time the two rounds of vertebrate genome duplication relative to the divergence of jawless and jawed vertebrates (Simakov et al. 2020).

6.5.2 Dot plots

The simplest way to compare the order of two sequences is the **dot plot**, introduced by Gibbs and McIntyre in 1970 (Gibbs and Mcintyre 1970). One sequence is placed along the horizontal axis and the other along the vertical axis, and a dot is drawn wherever the two are similar. With single residues the plot is dominated by random matches, so a dot is usually drawn only if a window of w residues contains at least t matches, or if its alignment score exceeds a threshold. A larger window or a stricter threshold removes noise but hides short or divergent similarities. Dotter displays the scores as a greyscale and lets the user change the threshold interactively (Sonnhammer and Durbin 1995). For whole genomes, the dots are matches found by a whole-genome aligner such as MUMmer (Kurtz et al. 2004), or pairs of orthologous genes.

Rearrangements have characteristic signatures (Figure 6.1):

- a **collinear** region gives a diagonal line with positive slope;
- an **inversion** gives a segment with the opposite slope, between two breakpoints, because the matching sequence is on the reverse strand. Both strands must therefore be compared for DNA;
- an **insertion or deletion** shifts the diagonal horizontally or vertically, leaving a gap;
- a **transposition** moves a segment of the diagonal to another position;
- a **duplication** gives two or more matches for one interval, and repeats give parallel lines or grids. Comparing a sequence with itself shows internal repeats as lines parallel to the main diagonal, and inverted repeats as lines perpendicular to it;
- with several chromosomes, a **translocation** moves a block to another chromosome, and **fusions** and **fissions** join or split blocks.

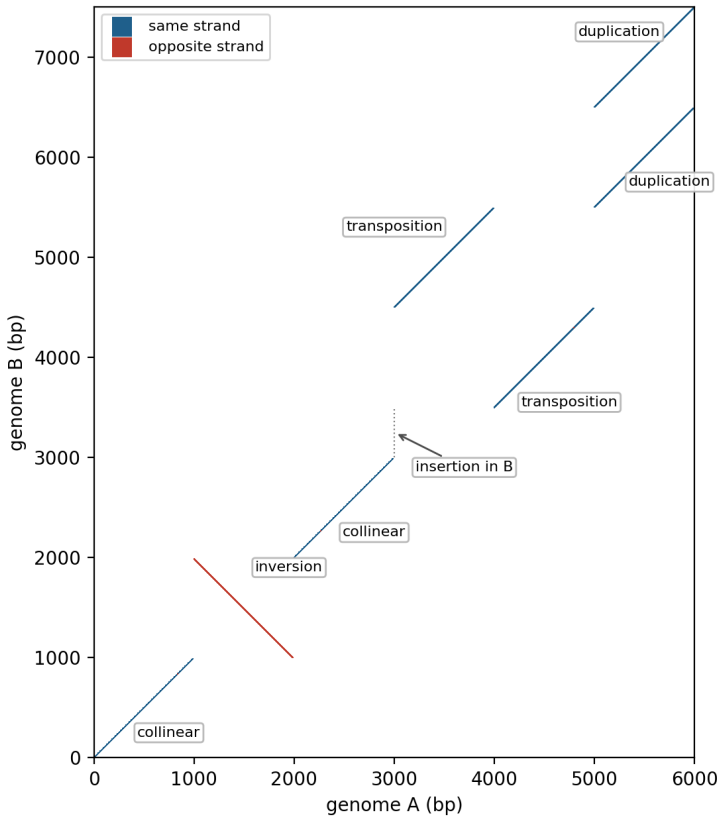


Figure 6.1: A dot plot of two synthetic genomes. Genome B was made from genome A (six blocks of 1 kb) by inverting the second block, inserting 500 bp of new sequence, swapping the order of the fourth and fifth blocks, and duplicating the last block. Dots mark identical 12-mers on the same strand (blue) or the opposite strand (red).

A dot plot proposes events but does not prove them. It shows similarity, not ancestry, so a clean diagonal can join paralogous repeats. It does not show which genome changed, which requires an outgroup. And assembly errors, such as misjoined contigs or a contig in the wrong orientation, produce exactly the same patterns as real rearrangements. A breakpoint should therefore be confirmed with reads or long molecules spanning it (Section 5.3) and, ideally, in an independent assembly.

6.5.3 Rearrangement distances

For computation, a genome is represented as an ordered list of genes or conserved blocks, each with a sign for its orientation. An inversion reverses the order of a segment and flips the signs. Inverting genes 2 to 4 of

$$(+1, +2, +3, +4, +5)$$

gives

$$(+1, -4, -3, -2, +5).$$

A **breakpoint** is an adjacency present in one genome but not in the other; this inversion created two, between $+1$ and -4 and between -2 and $+5$. The **breakpoint distance** counts them. The **reversal distance** is the smallest number of inversions that transforms one genome into the other. For signed permutations it can be computed in polynomial time, as Hannenhalli and Pevzner showed, using the example of the mitochondrial genomes of cabbage and turnip, which differ mainly by a few inversions (Hannenhalli and Pevzner 1999). The *double-cut-and-join* (DCJ) model treats inversions, translocations, fusions and fissions as special cases of one operation, which cuts two adjacencies and rejoins the ends in a new way, and gives particularly simple distances (Yancopoulos et al. 2005).

These distances are parsimony estimates, with the usual limitations: the true number of events can be larger than the minimum, repeated events at the same place overwrite each other, and duplications and gene losses violate the assumption that each gene occurs once in each genome. Whether breakpoints are scattered at random, as in the *random breakage model* of Nadeau and Taylor, was questioned when Pevzner and Tesler found that many human–mouse breakpoints have been reused, suggesting fragile regions that break repeatedly (Pevzner and Tesler 2003). Comparative maps of many mammals support breakpoint reuse and show that rates of rearrangement vary between lineages and over time (Murphy et al. 2005).

6.5.4 Detecting synteny at genome scale

Synteny is detected by finding **anchors**, usually pairs of homologous or orthologous genes, or local matches from a whole-genome alignment, and chaining them into blocks. The rules decide the result: how many anchors a block needs, how large a gap is allowed between them, whether order and orientation must be conserved, and how tandem duplicates are handled. MCScanX (Wang et al. 2012) is widely used for plant and animal genomes, and Mauve aligns bacterial genomes into collinear blocks while allowing rearrangements (Darling et al. 2004). Because a few genes in the same order can occur by chance, the significance of a block can be tested by comparison with randomised gene orders. Blocks can only be as long as the assembly allows, so fragmented assemblies give apparent breaks in synteny, and coordinates from different assembly versions should not be combined.

See also: further explanation in Section D.6.7; related Wikipedia articles in Section E.6.7.

Sources

Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Gene Order, Synteny, and Rearrangements" (CC BY 4.0), used with permission.

6.6 Genomes in populations

A single genome sequence is a sample from a population. Sequencing many individuals of the same species shows how much genomes vary, why they have the architecture they have, and, for our own species, how we are related to extinct humans.

6.6.1 Pan-genomes

When several strains of the same bacterial species were sequenced, they turned out to differ dramatically in gene content. Tettelin and colleagues, comparing eight genomes of *Streptococcus agalactiae*, introduced the **pan-genome**: the set of all genes found in a species, divided into a **core genome** present in all strains and an **accessory** (dispensable) genome present in only some, including strain-specific genes (Tettelin et al. 2005; Medini et al. 2005). In *E. coli*, each genome has about 4,500–5,000 genes, but only about 2,000 are shared by all strains, and 20 genomes together contain about 18,000 gene families (Touchon et al. 2009). Every new genome adds new genes. Such a pan-genome is called **open**, as opposed to a **closed** pan-genome, which stops growing after a few genomes have been sequenced. Open pan-genomes are typical of species with large populations that live in varied environments and exchange genes frequently, while specialised intracellular bacteria have small, closed pan-genomes.

The accessory genome is acquired by **horizontal gene transfer** and lost by deletion (Section 3.7). It contains many genes for niche-specific functions, such as antibiotic resistance, virulence and the use of particular nutrients, but also a large number of genes of unknown function, mobile elements and phages. Whether pan-genomes are mainly *adaptive*, with accessory genes maintained because they are useful in some environments, or mainly *neutral*, a by-product of transfer and drift in large populations, is debated (McInerney et al. 2017). Core genes are used to build species phylogenies, while accessory genes are used to type strains and to find genes associated with traits such as virulence.

6.6.2 From one reference genome to a pangenome

The concept has spread to eukaryotes, where it has changed what a reference genome is. Plant species such as maize show extensive presence–absence variation of genes between individuals, so no single individual contains all the genes of the species. For humans, the reference assembly GRCh38 is a single linear sequence, a mosaic of DNA from a few donors, about 70% of it from one individual, with alternative sequences for only a few highly variable regions such as the MHC. T2T-CHM13 is complete but represents one haploid genome (Section 6.3). Neither represents human diversity. Every person differs from the reference at millions of sites and carries more than 20,000 structural variants, some alleles common in populations under-represented among the donors are missing, and sequence absent from the reference cannot be analysed at all. Mapping reads to a single linear reference therefore causes *reference bias* (Section 5.4).

The Human Pangenome Reference Consortium instead represents the species as a collection of phased, near-complete assemblies of individuals from diverse populations, 47 in its first release, combined into a *pangenome graph* in which shared sequence is stored once and variants form alternative paths. Compared with GRCh38, mapping to the graph reduced errors in small-variant calls by about a third and

roughly doubled the number of structural variants detected per haplotype (Liao et al. 2023). GRCh38 remains important: it is the coordinate system of most annotations, variant databases and clinical pipelines, and graph-based coordinates are still being standardised. The conceptual change is that a linear reference is a convenient coordinate system, not a description of the species.

The word pangenome thus has two meanings (Figure 6.2). The *gene-content* pangenome of bacteria is a table of gene families against genomes; the *sequence* pangenome of humans and plants is a graph of alternative sequences. Both describe a particular sample. Whether a gene is "core" depends on the genomes included, the clustering of genes into families and the presence threshold (often 95% or 99% of genomes rather than all), and incomplete assemblies or missed gene calls make core genes look accessory. Whether a pan-genome is open or closed is likewise an estimate from the sample, not a fixed property of the species.

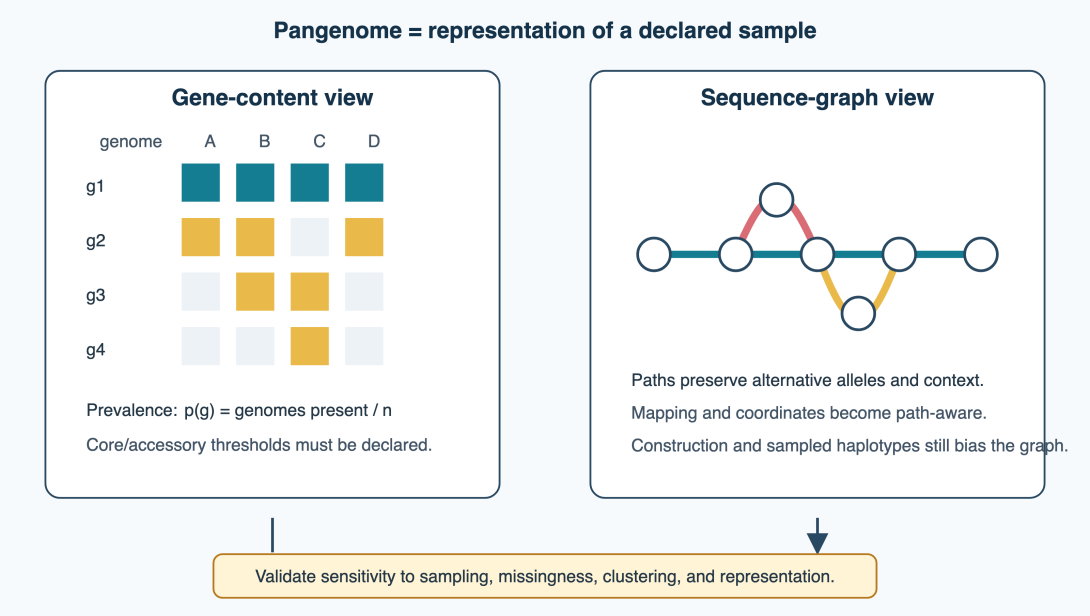


Figure 6.2: Two views of a pangenome. The gene-content view records which gene families are present in which genomes, and divides them into core and accessory families by a declared prevalence threshold. The sequence-graph view stores shared sequence once and represents variants as alternative paths, so that the haplotypes and their context are kept. Both depend on which genomes were sampled and how the data were processed. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

6.6.3 Population-genetic explanations of genome complexity

Why do prokaryotic genomes remain compact, while multicellular eukaryotes accumulate introns, transposable elements and non-coding DNA? Michael Lynch and John Conery proposed that the answer lies in population genetics rather than in adaptation (Lynch and Conery 2003). The argument builds on the nearly neutral theory (Section 3.3):

1. Most of the "extra" DNA of eukaryotes is slightly harmful. An intron, for example, adds splice sites that can mutate and cause mis-splicing, and a transposable element can disrupt genes when it moves. Lynch called this the **mutational hazard**: each additional element is a target for harmful mutations, so its selective disadvantage s is roughly proportional to the mutation rate u per site times the number of sites that must be maintained.
2. Selection can remove such elements only if s is larger than about $1/N_e$, where N_e is the effective population size.

3. Prokaryotes have very large effective population sizes (often above 10^8), so even slightly harmful DNA is removed and the genome remains streamlined. Large multicellular eukaryotes have much smaller ones (around 10^4–10^5), so drift lets introns, mobile elements and duplicate genes accumulate passively.

In this view, genome complexity is not an adaptation but a consequence of weak selection, which may later have been exploited. For example, introns made alternative splicing possible, and duplicated genes provided material for new functions. Lynch supported the theory with correlations between genome features and estimates of $N_e u$ across species, and later extended it to explain the evolution of mutation rates (the drift-barrier hypothesis; Section 3.3) (Lynch et al. 2016).

The theory is influential but debated. The correlations are confounded by phylogeny and by other variables, such as cell size and metabolic rate, and estimates of N_e are uncertain. In prokaryotes, the size of the genome seems to be set by a balance between a deletional bias and selection for acquired genes that are useful, rather than by drift alone (Sela et al. 2016). Two cases fit the theory well: endosymbiotic bacteria, whose tiny populations and bottlenecks at each host generation lead to the accumulation of harmful mutations and to genome decay (Nakabachi et al. 2006), and the streamlined genomes of the most abundant free-living bacteria, such as *Pelagibacter*, which live in enormous populations (Giovannoni et al. 2005).

6.6.4 Neanderthals and human evolution

The evolution of humans and the other great apes, as seen in phylogenetic trees, is described in Section 4.5. Genomics has added a detailed picture of the last few hundred thousand years.

Out of Africa. In 1987 a tree of mitochondrial DNA from people around the world traced all human mitochondria to a common ancestor in Africa about 200,000 years ago, the so-called "mitochondrial Eve" (Cann et al. 1987). Together with fossils, this supported the *out-of-Africa* model, in which modern humans arose in Africa and replaced other human populations elsewhere. Fossils from Jebel Irhoud in Morocco push the origin of *Homo sapiens* back to about 300,000 years ago (Hublin et al. 2017). Genome-wide data showed that genetic diversity is highest in Africa and decreases with distance from Africa, as expected from a series of founder events as humans spread across the world from about 60,000 years ago. Humans are genetically unusually homogeneous: two people differ at about one site in a thousand, and most of the variation is found within, rather than between, populations (The 1000 Genomes Project Consortium 2015).

Ancient DNA. DNA survives in bones for tens of thousands of years, but it is degraded into short fragments, carries characteristic damage (deamination of cytosine, which appears as C⟶T changes near the ends of the molecules), and is mixed with DNA from bacteria and modern humans. Svante Pääbo's group developed methods to deal with these problems. The first Neanderthal mitochondrial sequence, in 1997, fell outside the range of modern human variation, consistent with replacement (Krings et al. 1997). The draft Neanderthal nuclear genome in 2010 brought the surprise: Neanderthals share more variants with present-day non-Africans than with Africans, which shows that the ancestors of non-Africans **interbred with Neanderthals**, and that a few percent of the genome of present-day people outside Africa derives from Neanderthals (Green et al. 2010); current estimates are about 2%. Admixture was detected with statistics that count how often derived alleles are shared between pairs of populations in a four-taxon configuration, the D-statistic or ABBA–BABA test, a genome-scale cousin of the phylogenetic tests in Section 4.4.

The same year, DNA from a finger bone from Denisova Cave in Siberia revealed a previously unknown group of archaic humans, the **Denisovans**, known almost only from their genome (Reich et al. 2010). Present-day populations in Oceania carry several percent Denisovan ancestry. High-coverage genomes of a Denisovan (Meyer et al. 2012) and of a Neanderthal from the Altai mountains (Prüfer et al. 2013)

revealed several episodes of gene flow between archaic and modern groups, and the genome of "Denny", a girl with a Neanderthal mother and a Denisovan father, showed that the groups met and mixed directly (Slon et al. 2018). Pääbo was awarded the 2022 Nobel Prize in Physiology or Medicine for this work.

Consequences of admixture. Archaic DNA is not uniformly distributed in our genomes. It is depleted on the X chromosome and in genes expressed in the testes, which suggests that selection removed many archaic variants, possibly because they reduced fertility in hybrids (Sankararaman et al. 2014). Some archaic variants were instead favoured by selection (*adaptive introgression*): a Denisovan-like version of the gene *EPAS1* helps Tibetans live at high altitude (Huerta-Sánchez et al. 2014), and archaic variants in immune genes are common in many populations.

See also: further explanation in Section D.6.4; related Wikipedia articles in Section E.6.4.

Sources

Parts of this section are adapted from the Wikipedia articles Pan-genome, Genome size, Recent African origin of modern humans, Mitochondrial Eve, Neanderthal genome project, Interbreeding between archaic and modern humans and Denisovan (CC BY-SA 4.0, retrieved 3 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Genome Architecture, Representation, and Annotation" (CC BY 4.0), used with permission.

6.7 Repeats and the 3D genome

Most of a large eukaryotic genome consists of repeated sequences, and the genome is not a linear string but a molecule folded in three dimensions inside the nucleus. Both aspects matter for bioinformatics: repeats are the main problem in assembly and mapping (Chapter 5), and 3D folding brings regulatory elements into contact with the genes they control.

6.7.1 Transposable elements

Transposable elements (TEs) are DNA sequences that can move or copy themselves to new positions in the genome. Barbara McClintock discovered them in maize in the 1940s, from the unstable colour patterns of kernels (McClintock 1950), and received the Nobel Prize for the discovery in 1983. TEs are classified by their mechanism (Wells and Feschotte 2020):

- **Class I, retrotransposons**, move by "copy and paste": they are transcribed into RNA, reverse-transcribed into DNA and inserted at a new site, so each transposition adds a copy.
  - *LTR retrotransposons* and *endogenous retroviruses* have long terminal repeats, like retroviruses, and make up about 8% of the human genome.
  - *LINEs* (long interspersed nuclear elements) encode their own reverse transcriptase. **L1** makes up about 17% of the human genome, and about a hundred copies are still active.
  - *SINEs* (short interspersed nuclear elements) are short, non-autonomous elements that use the LINE machinery. The primate-specific **Alu** element, about 300 bp long, has more than a million copies and makes up about 10% of the human genome.
- **Class II, DNA transposons**, move by "cut and paste" using a transposase. They make up about 3% of the human genome and are no longer active in humans.

Altogether, recognisable TEs cover about half of the human genome, and more than 80% of the maize genome. Many older elements have decayed beyond recognition. Most insertions are neutral or slightly harmful: TEs can disrupt genes, and recombination between copies causes deletions and rearrangements. Genomes defend themselves with DNA methylation, repressive chromatin, small RNAs (piRNAs in the germline) and TE-silencing proteins. But TEs are also a source of novelty (Bourque et al. 2018): they provide promoters, enhancers and splice sites, and some have been *domesticated* into host genes. The RAG recombinase that drives the rearrangement of antibody genes derives from a DNA transposon, and the syncytin proteins essential for the placenta come from the envelope genes of endogenous retroviruses.

TEs are identified with libraries of consensus sequences or profile HMMs, such as Dfam (Storer et al. 2021), searched with tools such as RepeatMasker, and new families are found *de novo* from repeated sequence in a new genome. Masking repeats is the first step in gene annotation (Section 5.6).

6.7.2 Tandem repeats and segmental duplications

Tandem repeats are copies placed head to tail. **Microsatellites** (short tandem repeats, with units of 1–6 bp) mutate rapidly by replication slippage. They are used in forensics and population genetics, and their expansion causes diseases such as Huntington's disease (CAG repeats). **Satellite DNA** consists of long arrays of longer units. Human centromeres are built from megabase arrays of the 171 bp *alpha satellite*, organised in higher-order repeats. They were the largest gaps in the human reference until the complete T2T genome made it possible to map them (Nurk et al. 2022; Altemose et al. 2022).

Segmental duplications are copies of more than 1 kb with more than 90% identity, often at different chromosomal positions. They make up about 5–7% of the human genome, are enriched near centromeres and telomeres, and are particularly frequent in great apes (Bailey et al. 2002). They are a source of new genes, including several human-specific genes involved in brain development, but recombination between copies also causes recurrent deletions and duplications associated with disease. Because the copies are nearly identical, segmental duplications were the regions most often collapsed or misassembled in reference genomes (Section 5.3).

6.7.3 How structural variants arise

Structural variants (SVs), such as deletions, duplications, inversions and translocations, arise when DNA breaks or replication goes wrong and the pieces are joined incorrectly. The main mechanisms leave different signatures at the breakpoint junctions, so the junction sequence, which long reads often resolve (Section 5.4), tells how an SV was formed (Hastings, Lupski, et al. 2009; Carvalho and Lupski 2016).

- **Non-allelic homologous recombination** (NAHR) is recombination between two similar copies at different positions, such as segmental duplications or, at a smaller scale, Alu and L1 elements. Between copies in direct orientation on the same chromosome it deletes or duplicates the region between them, as reciprocal products; between inverted copies it inverts the region; between chromosomes it gives translocations. Because the duplications define the breakpoints, NAHR produces *recurrent* rearrangements with the same size and breakpoints in unrelated individuals (Lupski 1998).
- **Non-homologous end joining** (NHEJ) repairs double-strand breaks by joining the ends directly, with no or only a few bases of homology (*microhomology*), often with small deletions or insertions at the junction (Lieber 2010). The resulting SVs are *non-recurrent*: each has its own breakpoints.
- **Replication-based mechanisms.** When a replication fork stalls or collapses, the end of the new strand can invade another template nearby, using microhomology, copy some sequence, and

switch again. This *fork stalling and template switching* (FoSTeS) (Lee et al. 2007), or *microhomology-mediated break-induced replication* (MMBIR) (Hastings, Ira, et al. 2009), explains non-recurrent duplications with complex junctions, short inserted sequences copied from elsewhere, and patterns such as duplication–triplication–duplication.

- **Chromothripsis** ("chromosome shattering") is a single catastrophic event in which one or a few chromosomes, or chromosome arms, break into many pieces that are rejoined in random order and orientation, with some pieces lost (Stephens et al. 2011). Its signature is tens to hundreds of breakpoints clustered in a region whose copy number oscillates between two states. It was discovered in cancer genomes, where it occurs in a few per cent of tumours overall and more often in some types, and is thought to follow when a chromosome is trapped in a micronucleus or a chromatin bridge. It also occurs in the germline.

Retrotransposition of L1 and Alu elements and of processed mRNAs (see above and Section 6.8) adds a further class of insertions.

Genomic disorders are diseases caused by such rearrangements (Lupski 1998). Many recurrent ones are flanked by segmental duplications and caused by NAHR. In 17p12, a 1.4 Mb duplication that includes *PMP22* causes Charcot–Marie–Tooth disease type 1A, and the reciprocal deletion causes a different neuropathy (HNPP). The 22q11.2 deletion syndrome and Williams–Beuren syndrome (7q11.23) arise in the same way. The phenotype follows from **dosage sensitivity**: some genes cause disease when one copy is lost (*haploinsufficiency*) or when an extra copy is present (*triplosensitivity*), often genes whose products work in balanced complexes or networks (Section 3.4). Most genes tolerate a change in copy number, and population catalogues of SVs, such as gnomAD-SV (Collins et al. 2020), help to separate deletions and duplications that are common, and therefore probably harmless, from rare ones in constrained genes. How SVs are detected from sequencing data is described in Section 5.4.

6.7.4 Chromatin and 3D genome organisation

The DNA in a human cell is about 2 m long and must fit into a nucleus a few micrometres across. It is wrapped around histone octamers in **nucleosomes**, about 147 bp each, which are further compacted into chromatin. Loosely packed, active **euchromatin** is distinguished from dense, mostly silent **heterochromatin**, with differences in histone modifications and DNA methylation, the main carriers of *epigenetic* information. Each chromosome occupies its own *territory* in the nucleus.

The 3D organisation inside these territories was revealed by **chromosome conformation capture** methods. In 3C, chromatin is cross-linked, cut with a restriction enzyme and religated, so that pieces of DNA that were cross-linked to each other, usually because they were close in space, are joined and can be detected (Dekker et al. 2002). **Hi-C** sequences all such ligation junctions, giving a genome-wide matrix of contact frequencies (Lieberman-Aiden et al. 2009).

A contact frequency is not a measurement of physical distance. It counts how often two regions were cross-linked and ligated together in a population of millions of cells. A high frequency can mean that the regions are close in most cells or in contact in only a few, and single-cell Hi-C and microscopy show that the structure varies greatly from cell to cell. Frequencies also depend on the efficiency of cross-linking and ligation, on the density of restriction sites, on GC content and on mappability. They are dominated by genomic distance: as in any polymer, sites that are close along the chromosome are in contact far more often than distant ones. Contact maps are therefore normalised and compared with this expected decay before structures are inferred. With these caveats, Hi-C data showed several levels of organisation:

- **Compartments.** The genome is divided into an active *A* compartment and an inactive *B* compartment, which segregate in the nucleus. They are found with principal component analysis of the contact matrix (Lieberman-Aiden et al. 2009).

- **Topologically associating domains** (TADs). At the sub-megabase scale, the genome consists of domains within which contacts are frequent and across which they are rare. TADs are largely shared between cell types, and many are conserved between species (Dixon et al. 2012; Nora et al. 2012). Like other Hi-C structures, they describe the average of a cell population.
- **Loops.** High-resolution maps showed about 10,000 loops in the human genome, mostly between pairs of binding sites for the protein CTCF in convergent orientation (Rao et al. 2014). They are explained by **loop extrusion**: the cohesin complex extrudes a loop of DNA until it meets CTCF bound in the right orientation (Fudenberg et al. 2016).

TADs and loops constrain which enhancers can contact which promoters. Deletions or inversions that remove a TAD boundary can let an enhancer activate the wrong gene, which causes several developmental disorders. Analysing Hi-C data is itself a computational field: contact matrices must be normalised for the biases above, and domains and loops are called with dedicated algorithms. Hi-C is also used to scaffold genome assemblies, since contacts are much more frequent within than between chromosomes (Section 5.3).

See also: further explanation in Section D.6.5; related Wikipedia articles in Section E.6.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Transposable element, Retrotransposon, Alu element, Tandem repeat, Segmental duplication, Chromatin, Chromosome conformation capture and Topologically associating domain (CC BY-SA 4.0, retrieved 3 October 2026).

6.8 Pseudogenes, new genes and small proteins

A gene catalogue (Section 6.3) is not fixed. Genes die and become pseudogenes, new genes are born, and some real genes are missing from the catalogues because they are too short to be found by the usual methods. All three matter in practice: pseudogenes confuse mapping and quantification, young genes have no homologues to annotate them from, and small proteins fall below the length cut-offs of gene prediction.

6.8.1 Pseudogenes

A **pseudogene** is a copy of a gene that has lost its ability to encode a functional protein, through stop codons, frameshifts, loss of a promoter or truncation. The human genome contains more than 10,000 recognisable pseudogenes, about as many as protein-coding genes, and the GENCODE annotation distinguishes three main types (Pei et al. 2012):

- **Processed pseudogenes** (retrocopies) arise when an mRNA is reverse-transcribed by the L1 machinery and inserted elsewhere (Section 6.7). They lack introns and promoters, often carry the remains of a poly(A) tail, and are the majority in mammals. Highly expressed genes, such as those for ribosomal proteins and GAPDH, have dozens of them.
- **Duplicated (unprocessed) pseudogenes** arise from a DNA-level duplication (Section 3.4) followed by loss of function of one copy. They keep the exon–intron structure and often lie near their parent.
- **Unitary pseudogenes** are former single-copy genes that were disabled in place, without a functional copy remaining. A classic example is *GULO*, needed for vitamin C synthesis, which is a pseudogene in primates (and in guinea pigs); humans must therefore obtain vitamin C from their diet.

Pseudogenes are a practical problem in bioinformatics. A processed pseudogene looks like the spliced mRNA of its parent, so RNA-seq reads from the parent can align to it without splitting, and reads from the pseudogene, if it is transcribed, can be assigned to the parent. Expression estimates for such genes depend on how multi-mapping reads are handled (Section 7.3). In DNA sequencing, a variant in a pseudogene can be mistaken for a variant in the gene, and true variants can be lost because reads map with low quality to both copies (Section 5.4). Clinical genes with close pseudogenes, such as *PMS2*, *CYP21A2* and *GBA1*, need special methods, and gene conversion between gene and pseudogene can move pathogenic variants into the functional copy. Retrocopies that are present in only some individuals appear as apparent intron deletions in DNA sequence. Annotation pipelines must also avoid calling pseudogenes as genes: a disrupted ORF can come from a sequencing error in a real gene, and a pseudogene can be mistaken for a gene if the disruption falls in a gap.

Most pseudogenes are thought to have no function and evolve neutrally, which makes them useful for estimating mutation rates. A minority are transcribed, and some of these have functions. The best-known proposal is that the transcript of *PTENP1* acts as a decoy that binds microRNAs targeting the tumour suppressor *PTEN* (Poliseno et al. 2010), although how general such "competing endogenous RNA" effects are is debated. Other pseudogenes produce small interfering RNAs, or are translated into short proteins (Cheetham et al. 2019). As for non-coding RNAs (Section 6.9), transcription alone does not show function.

6.8.2 How new genes are born

New genes arise in several ways (Long et al. 2003):

- **Duplication** of an existing gene, which is the main source of new genes in eukaryotes, followed by neo- or subfunctionalisation (Section 3.4).
- **Retroposition.** A retrocopy usually becomes a processed pseudogene, but occasionally it lands near a promoter or recruits one and becomes a functional *retrogene* (Kaessmann et al. 2009). In mammals, retrogenes copied from X-linked genes are often expressed in the testis, where they replace their parents when the X chromosome is silenced during male meiosis.
- **Gene fusion and fission** join parts of genes or split them, and rearrange protein domains (Section 3.5).
- **Horizontal gene transfer**, dominant in prokaryotes (Section 3.7).
- **Domestication** of transposable elements and viral genes, such as the RAG recombinase and the syncytins (Section 6.7).
- **De novo gene birth** from sequence that was not previously a gene.

The last mechanism was long considered improbable, but comparative genomics has found convincing cases in yeast, flies, plants and primates, including human genes with no counterpart in chimpanzee that arose from non-coding DNA (Knowles and McLysaght 2009; Van Oss and Carvunis 2019). Non-coding DNA contains many short ORFs, and much of it is transcribed (Section 6.9), so the raw material is plentiful. Using ribosome profiling in yeast, Carvunis and colleagues found that many short, non-conserved ORFs are translated, and that their properties, such as length, codon usage and expression, lie between those of random non-genic ORFs and those of established genes (Carvunis et al. 2012). They proposed a continuum of **proto-genes**, from which a few gain a function, are retained by selection and become genes. In most cases, a transcript gains an ORF, or an ORF gains transcription, and the order in which these happen differs between cases.

De novo genes are a subset of **orphan genes**, or *taxonomically restricted genes*: genes with no detectable homologues outside a lineage. Every newly sequenced genome has many of them, often 10–30% of the genes (Tautz and Domazet-Lošo 2011). Not every orphan is new, however. A gene that evolves fast or is short can lose detectable similarity to its homologues, so that sequence searches miss them (1).

Simulations suggest that such *homology detection failure* explains many apparent orphans (Weisman et al. 2020), while analyses that use conserved gene order to find the corresponding region in related genomes show that only a minority of orphans can be explained in this way (Vakirlis et al. 2020). A convincing de novo gene is therefore shown by finding the orthologous region in related species, demonstrating that it is non-coding there, and identifying the mutations that created the ORF or its expression. Methods that date genes by the most distant species in which a homologue is found (*phylostratigraphy*) are biased by the same problem, since short and fast genes appear younger than they are.

6.8.3 Small proteins and smORFs

Small proteins, or **microproteins**, are usually defined as those shorter than about 100 amino acids, encoded by **small ORFs** (smORFs). They were systematically missed by genome annotation (Storz et al. 2014; Couso and Patraquim 2017). In random sequence an ORF of 50 codons is common (Section 5.6), so early annotations set a minimum length, typically 100 codons, to keep the false positives down. Short genes also give weak statistical signals for gene finders, have few or no homologues in the databases, and give few peptides in proteomics. Some well-known small proteins were nevertheless found early, by biochemistry: phospholamban and sarcolipin, which regulate the calcium pump SERCA in muscle, are 52 and 31 residues long.

Small ORFs come in several classes (Couso and Patraquim 2017): short genes on their own transcripts; *upstream ORFs* (uORFs) in the 5' UTRs of mRNAs, which mostly regulate the translation of the main ORF; ORFs in transcripts annotated as long non-coding RNAs; and *alternative ORFs* that overlap known coding sequence in another frame. Three kinds of evidence are used to find them:

- **Ribosome profiling** (Ribo-seq) sequences the fragments of mRNA protected by ribosomes, and so shows which regions are translated, at single-nucleotide resolution (Ingolia et al. 2009). Translated ORFs show a three-nucleotide periodicity of the footprints. Ribo-seq has revealed thousands of translated smORFs in human cells and tissues (Chothani et al. 2022), but ribosome binding does not prove that a stable protein is made.
- **Proteomics and peptidomics.** Mass spectrometry can confirm the protein, but small proteins are often of low abundance and give one or two tryptic peptides. Enrichment of small proteins, searches without digestion and immunopeptidomics, which detects peptides presented by HLA molecules, increase sensitivity (Slavoff et al. 2012). Searches against databases of all possible ORFs are proteogenomic searches with a large search space and need strict error control (Section 8.3).
- **Evolutionary conservation.** Codon substitution patterns in alignments of related genomes show whether an ORF evolves like protein-coding sequence. PhyloCSF (Lin et al. 2011) compares the likelihoods of the observed codon substitutions under coding and non-coding models. It works for conserved smORFs but, by design, not for young ones.

Several small proteins with important functions were found in transcripts annotated as non-coding. *Myoregulin* (46 residues) (Anderson et al. 2015) and *DWORF* (34 residues) (Nelson et al. 2016) are single transmembrane helices that inhibit or activate SERCA in muscle. Other microproteins act in mitochondria, for example in the respiratory chain, and the peptides of the *polished rice* (*tarsal-less*) gene in *Drosophila*, 11–32 residues long, control development (Kondo et al. 2010). In bacteria, small proteins regulate membrane transporters, kinases and cell division (Storz et al. 2014), and the human microbiome encodes thousands of families of small proteins (Sberro et al. 2019).

How many of the translated smORFs encode functional proteins is the open question. A CRISPR screen found that disrupting some non-canonical ORFs affected cell growth (Chen et al. 2020), but most translated smORFs are short, young and not conserved, and many may be translational noise or regulatory uORFs. To bring them into the reference annotation, GENCODE has started to include a

curated set of about 7,000 Ribo-seq ORFs, annotated with the strength of their evidence (Mudge et al. 2022).

Structure prediction (Section 9.8) is increasingly applied to small proteins, to find out whether they fold and with what they interact. It is harder than for ordinary proteins. Many microproteins are single helices, often transmembrane, or are disordered, and young ones have few or no homologues, so the multiple sequence alignments that AlphaFold relies on are shallow. Predictions therefore often have low confidence, and a confident model of a short peptide is not by itself evidence that it folds. Language-model predictors that work on single sequences (Section 9.9) avoid the alignment problem, but not the underlying lack of information. The most promising use is to predict complexes between microproteins and candidate partners, as for the SERCA regulators. Interfaces of short peptides are, however, predicted less reliably than folds, and screens of many pairs produce false positives that need experimental testing (Section 9.8).

See also: further explanation in Section D.6.8; related Wikipedia articles in Section E.6.8.

> **i** Sources
>
> This section was written for this book from the primary literature cited.

6.9 RNA genes and regulatory elements

Protein-coding exons are a small part of a eukaryotic genome. Much of the rest is transcribed, and some of it encodes functional RNAs or regulates the expression of genes. How much of the non-coding genome is functional is one of the most debated questions in genomics.

6.9.1 Non-coding RNAs

Genes whose product is an RNA rather than a protein are called non-coding RNA (ncRNA) genes. The search for them, and the prediction of their structures, is described in Section 1.7. The main classes are:

- **Housekeeping RNAs**, present in all cells and essential for basic functions: ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) in translation, small nuclear RNAs (snRNAs) in the spliceosome, small nucleolar RNAs (snoRNAs) that guide the modification of other RNAs, and the RNA components of ribonuclease P, the signal recognition particle and telomerase. Several of them are catalytic, a possible relic of the RNA world (Section 3.2).
- **MicroRNAs** (miRNAs), about 22 nucleotides long, regulate gene expression by base pairing with the 3' untranslated regions of mRNAs, which leads to their degradation or reduced translation. The first, *lin-4*, was found in *C. elegans* in 1993 (Lee et al. 1993). The human genome encodes hundreds of conserved miRNAs, which together regulate a large fraction of protein-coding genes (Bartel 2018). Victor Ambros and Gary Ruvkun received the 2024 Nobel Prize for their discovery.
- **Small interfering RNAs** (siRNAs) and **piRNAs** silence viruses and transposable elements; piRNAs protect the germline (Section 6.7).
- **Long non-coding RNAs** (lncRNAs), longer than 200 nucleotides and not translated. Tens of thousands are annotated in the human genome. A few are well studied, such as *XIST*, which silences one X chromosome in females. Most are expressed at low levels, are poorly conserved and have no known function, and it is debated how many are functional and how many are transcriptional noise (Mattick et al. 2023; Palazzo and Lee 2015).
- In bacteria, small regulatory RNAs, *riboswitches* (structures in mRNAs that bind metabolites and control expression) and the CRISPR RNAs of bacterial immunity.

6.9.2 Regulatory elements

Gene expression is controlled by DNA sequences that bind transcription factors and other proteins:

- **Promoters**, at the start of a gene, where RNA polymerase and the general transcription factors assemble. Many human promoters lie in *CpG islands*, regions that escaped the general depletion of CpG dinucleotides caused by the mutation of methylated cytosines.
- **Enhancers**, which increase transcription and can lie up to a megabase away from their target gene, in introns or in intergenic regions. They act by contacting the promoter through chromatin loops (Section 6.7).
- **Silencers**, which repress transcription, and **insulators**, which block the action of enhancers across boundaries, often by binding CTCF.

Regulatory elements are short and degenerate, so they are hard to find from sequence alone. They are mapped experimentally from their biochemical signatures: open chromatin (DNase-seq, ATAC-seq), specific histone modifications (such as H3K27ac at active enhancers and H3K4me3 at active promoters), and transcription-factor binding (ChIP-seq). Evolutionary conservation provides an independent signal. Some non-coding elements, the *ultraconserved elements*, are identical over hundreds of bases between human, mouse and rat, more conserved than most protein-coding genes. The ENCODE project has catalogued close to a million candidate regulatory elements in the human genome (The ENCODE Project Consortium 2020). The analysis of the corresponding expression data is described in Chapter 7.

6.9.3 How much of the genome is functional?

The term *junk DNA* was popularised by Susumu Ohno in 1972 for DNA, such as the decayed copies of duplicated genes, that had accumulated without a function. The C-value paradox (Section 6.3) was a strong argument for it: if an onion needs five times more DNA than a human, or one species of amphibian much more than a close relative, most of that DNA can hardly be essential.

Junk DNA means DNA with no known *selected* function: mutating or deleting it would not affect the fitness of the organism. It does not mean DNA without biochemical activity. Junk DNA may be transcribed, bound by proteins or carry chromatin marks, because these activities also occur by chance. Conversely, transcription or protein binding alone does not show that a sequence has a function for the organism. That requires evidence of evolutionary constraint or of a phenotypic effect when the sequence is mutated or deleted.

In 2012 the ENCODE project reported that about 80% of the human genome had a "biochemical function", meaning that it was transcribed, bound by proteins or marked by specific chromatin features in at least one cell type (The ENCODE Project Consortium 2012). The press reported the death of junk DNA. Evolutionary biologists responded sharply (Eddy 2012; Graur et al. 2013; Doolittle 2013). They argued that ENCODE confused two meanings of function: *causal-role* function, an activity that can be measured, and *selected-effect* function, an activity that is maintained by natural selection. Much transcription and protein binding may be biochemical noise, and DNA can be active without being needed. ENCODE scientists later acknowledged that the different definitions give very different numbers (Kellis et al. 2014).

Evolutionary estimates of the functional fraction measure how much of the genome is under purifying selection, from the deficit of substitutions compared with neutral expectations. Comparisons of mammalian genomes give estimates of about 8% for the human genome (Rands et al. 2014), and the Zoonomia alignment of 240 mammals identified at least 3% of human bases as significantly constrained and estimated that around 10% are under constraint (Christmas et al. 2023). These estimates miss recently evolved and rapidly turned-over functional elements, but they are hard to reconcile with 80%. A population-genetic argument points in the same direction: each human inherits about 60–70 new

mutations (Section 3.3), and if most of the genome were sensitive to mutation the load of harmful mutations would be very high. Taken together, these arguments suggest that perhaps 10–15% of the human genome is functional in the selected-effect sense, but the question is far from settled, and it is a typical exam topic because it connects genome biology, evolution and the interpretation of large-scale data.

See also: further explanation in Section D.6.6; related Wikipedia articles in Section E.6.6.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Non-coding RNA, MicroRNA, Long non-coding RNA, Enhancer (genetics), Non-coding DNA, Junk DNA and ENCODE (CC BY-SA 4.0, retrieved 3 October 2026).

6.10 Key concepts

A checklist for revision.

- **Background:** DNA and RNA structure and base pairing; replication, transcription, RNA processing and translation; the genetic code; reverse transcription; levels of gene regulation; prokaryotic and eukaryotic cells; model organisms
- **Genome projects:** the first genomes (*H. influenzae*, yeast, *E. coli*, *C. elegans*, *Drosophila*, *Arabidopsis*, human, mouse); clone-by-clone vs. whole-genome shotgun; the public–Celera race; Bermuda principles; what "finished" means and the complete T2T genome; assembly vs. annotation and their versions
- **What is a gene?** Mendelian, molecular and post-ENCODE definitions (a union of sequences encoding related functional products); why the human protein-coding count fell from ~100,000 guesses to ~19,000–20,000; GENCODE/Ensembl, RefSeq, CHESS and MANE, and where they disagree (uncertain coding status, readthrough, duplicated families, isoforms)
- **Genome size and gene number:** C-value; range across bacteria, archaea and eukaryotes; smallest and largest genomes; gene density; gene number proportional to genome size in prokaryotes but not in eukaryotes; genome size dominated by repeats, not gene number; polyploidy and monoploid genome size; the C-value paradox and C-value enigma; the G-value paradox (gene number does not measure complexity)
- **Prokaryotic genome organisation:** circular chromosomes and plasmids; single origin and replichores; GC skew; strand bias and gene position; operons; compactness and deletional bias; genomic islands; archaea as a mix
- **Eukaryotic genome organisation:** linear chromosomes, telomeres and telomerase, centromeres; multiple origins; introns, splicing and the discovery of split genes; introns early vs. late; alternative splicing; large intergenic regions and distal regulation
- **Genome duplications:** paralogous synteny and 2:1 patterns; yeast WGD, vertebrate 2R, teleost and plant polyploidy; allo- and autopolyploids, homoeologues, and why polyploid and hybrid genomes are hard to assemble and to analyse for orthology
- **Organelle genomes:** endosymbiotic theory; origins of mitochondria and plastids; human mtDNA; variation in organelle genomes; endosymbiotic gene transfer and NUMTs; genome reduction in endosymbionts
- **Viral genomes:** DNA or RNA, single- or double-stranded, segmented genomes; Baltimore classes vs. taxonomy; hallmark genes and realms; genome sizes from circoviruses to pandoraviruses; overlapping genes (φX174), polyproteins, frameshifting; high mutation rates of RNA viruses; consequences for gene prediction and phylogeny

- **Synteny and rearrangements:** synteny vs. collinearity; conserved segments (Nadeau and Taylor) and differing rates of rearrangement; uses of synteny (orthology, missing genes, WGD, ancestral linkage groups); dot plots, window and threshold, and the signatures of inversions, indels, transpositions, duplications and translocations; assembly errors that mimic breakpoints; signed permutations, breakpoint and reversal distances, DCJ; random breakage vs. breakpoint reuse; anchors and chaining (MCScanX, Mauve)
- **Pan-genomes:** core and accessory genomes; open vs. closed pan-genomes; HGT; adaptive vs. neutral explanations; gene-content vs. sequence pangenomes; core defined by a sampling-dependent threshold; eukaryotic and human pangenomes; linear reference (GRCh38, CHM13) vs. pangenome graph; reference bias
- **Population genetics of genome complexity:** Lynch and Conery; the mutational hazard; N_e and the efficiency of selection; criticism; endosymbionts and streamlined free-living bacteria
- **Human evolution:** mitochondrial Eve and out of Africa; low human diversity; ancient DNA and its damage; Neanderthal and Denisovan genomes; admixture and the D-statistic; introgression deserts and adaptive introgression
- **Repeats:** classes of transposable elements (LTR, LINE, SINE, DNA transposons); L1 and Alu; host defences and domestication; repeat masking; microsatellites, satellites and centromeres; segmental duplications
- **Mechanisms of structural variation:** NAHR between segmental duplications (recurrent, reciprocal deletions and duplications); NHEJ and microhomology; replication-based FoSTeS/MMBIR and complex junctions; chromothripsis; junction signatures; genomic disorders (CMT1A/HNPP, 22q11.2); haploinsufficiency, triplosensitivity and population SV catalogues
- **Pseudogenes and new genes:** processed, duplicated and unitary pseudogenes (GULO); how they confuse read mapping, variant calling and quantification; occasional functions (PTENP1) and the ceRNA debate; gene birth by duplication, retroposition (retrogenes), fusion/fission, HGT, TE domestication and de novo origin; proto-genes; orphan genes and homology detection failure; synteny-based evidence; phylostratigraphy bias
- **Small proteins:** microproteins and smORFs (<100 aa); why ORF-length cut-offs missed them; uORFs, lncRNA ORFs, alternative ORFs; Ribo-seq and its caveats; peptidomics and proteogenomic error control; PhyloCSF and conservation; myoregulin, DWORF, polished rice; GENCODE Ribo-seq ORFs; structure prediction of microproteins and its limits
- **3D genome:** nucleosomes and chromatin; eu- and heterochromatin; chromosome territories; 3C and Hi-C; contact frequency is not distance (population averaging, biases, polymer effects); A/B compartments, TADs and CTCF loops; loop extrusion
- **Non-coding RNAs:** housekeeping RNAs; miRNAs, siRNAs and piRNAs; lncRNAs and the debate about their function; bacterial sRNAs and riboswitches
- **Regulatory elements:** promoters and CpG islands; enhancers, silencers and insulators; how they are mapped; ultraconserved elements
- **The function debate:** junk DNA (no selected function, not no biochemical activity); ENCODE 2012 and its critics; causal-role vs. selected-effect function; estimates from evolutionary constraint; the mutational-load argument

6.11 Recommended reading

These papers are recommended for reading in full:

- International Human Genome Sequencing Consortium et al. (2001) — the human genome draft; what a genome looks like, and how it was obtained.
- Lynch and Conery (2003) — why prokaryotic and eukaryotic genomes differ; the population-genetic theory of genome complexity.

- McInerney et al. (2017) — why prokaryotes have pan-genomes; core and accessory genes, and adaptive vs. neutral explanations.
- Green et al. (2010) — the draft Neanderthal genome and the discovery of admixture with modern humans.
- Bourque et al. (2018) — ten things you should know about transposable elements.
- Eddy (2012) — the C-value paradox, junk DNA and ENCODE, in two pages.

Titles, abstracts, links and the full texts are in Section A.6; a longer list of further papers is in Section B.6.

6.12 Background reading (open access)

- **What is a genome? (YourGenome)** — An introduction for a general audience, from Wellcome Connecting Science, to genomes: DNA, genes, chromosomes and how genome sizes differ between organisms.
  https://www.yourgenome.org/theme/what-is-a-genome/
- **Genome (Wikipedia)** — In the fields of molecular biology and genetics, a genome is all the genetic information of an organism. It consists of nucleotide sequences of DNA.
  https://en.wikipedia.org/wiki/Genome
- **Genome size (Wikipedia)** — Genome size is the total amount of DNA contained within one copy of a single complete genome. It is typically measured in terms of mass in picograms or less frequently in daltons, or as the total number of nucleotide base pairs, usually in megabases. One picogram is equal to 978 megabases.
  https://en.wikipedia.org/wiki/Genome_size
- **C-value (Wikipedia)** — C-value is the amount, in picograms, of DNA contained within a haploid nucleus or one half the amount in a diploid somatic cell of a eukaryotic organism.
  https://en.wikipedia.org/wiki/C-value
- **Human genome (Wikipedia)** — The human genome is a complete set of DNA sequences for each of the 22 autosomes and the two distinct sex chromosomes. A small DNA molecule is found within individual mitochondria. These are usually treated separately as the nuclear genome and the mitochondrial genome.
  https://en.wikipedia.org/wiki/Human_genome
- **Human Genome Project (Wikipedia)** — The Human Genome Project (HGP) was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying, mapping and sequencing all of the genes of the human genome from both a physical and a functional standpoint.
  https://en.wikipedia.org/wiki/Human_Genome_Project
- **Operon (Wikipedia)** — In genetics, an operon is a functioning unit of DNA containing a cluster of genes under the control of a single promoter.
  https://en.wikipedia.org/wiki/Operon
- **Intron (Wikipedia)** — An intron is any nucleotide sequence within a gene that is not expressed or operative in the final RNA product. The word intron is derived from the term intragenic region, i.e., a region inside a gene. The term intron refers to both the DNA sequence within a gene and the corresponding RNA sequence in RNA transcripts.
  https://en.wikipedia.org/wiki/Intron
- **Mitochondrial DNA (Wikipedia)** — Mitochondrial DNA is the DNA located in the mitochondria organelles in a eukaryotic cell that converts chemical energy from organic compounds into adenosine triphosphate (ATP).
  https://en.wikipedia.org/wiki/Mitochondrial_DNA
- **Pan-genome (Wikipedia)** — In the fields of molecular biology and genetics, a pan-genome is the entire set of genes from all strains within a clade. More generally, it is the union of all the

genomes of a clade.
https://en.wikipedia.org/wiki/Pan-genome

- **Interbreeding between archaic and modern humans (Wikipedia)** — Interbreeding between archaic humans and anatomically modern humans took place during the Middle Paleolithic and early Upper Paleolithic.
https://en.wikipedia.org/wiki/Interbreeding_between_archaic_and_modern_humans
- **Transposable element (Wikipedia)** — Transposable elements (TEs), also known as transposons, jumping genes, or mobile genetic elements, are DNA sequences that can change their position, or translocate, within a genome.
https://en.wikipedia.org/wiki/Transposable_element
- **Topologically associating domain (Wikipedia)** — A topologically associating domain (TAD) is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD.
https://en.wikipedia.org/wiki/Topologically_associating_domain
- **Non-coding DNA (Wikipedia)** — Non-coding DNA (ncDNA) sequences are components of an organism's DNA that do not encode protein sequences. Some non-coding DNA is transcribed into functional non-coding RNA molecules.
https://en.wikipedia.org/wiki/Non-coding_DNA
- **Junk DNA (Wikipedia)** — Junk DNA is a DNA sequence that has no known biological function. Most organisms have some junk DNA in their genomes—mostly pseudogenes and fragments of transposons and endogenized viruses—but it is possible that some organisms have substantial amounts of junk DNA.
https://en.wikipedia.org/wiki/Junk_DNA

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

6.13 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

6.14 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Genome Architecture, Representation, and Annotation" and "Gene Order, Synteny, and Rearrangements" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

6.14.1 Comparing genome architectures without stereotypes

Compare the organisation of bacterial, archaeal, eukaryotic nuclear, organelle and viral genomes. Show how the differences determine which computational representations and analyses are appropriate.

Extensions, solution outline and common errors

Extensions:

1. Why need a bacterial chromosome be neither single nor circular?
2. Distinguish an operon from a set of neighbouring genes.
3. Which features of archaea resist a simple division into bacteria and eukaryotes?
4. Why can organelle genomes not be treated as miniature bacterial chromosomes?
5. Which properties of viral genomes change the strategy for assembly and annotation?

Solution outline:

- Compare genomes along independent axes: DNA or RNA and strandedness, topology, the number of molecules or segments, ploidy and copy number, packaging, coding strategy, repeats, and diversity within a sample.
- Treat the differences between domains as tendencies, not definitions. Bacterial and archaeal genomes are usually compact but can have several or linear replicons. Eukaryotic nuclear chromosomes are linear and packed in chromatin. Organelle genomes are reduced but vary in structure, genetic code and RNA processing. Viral genomes can be DNA or RNA, single- or double-stranded and segmented, and use overlapping genes, frameshifts and polyproteins.
- An operon is defined by a shared transcript. Adjacent genes on the same strand are only a prediction, because internal promoters and terminators and condition-specific boundaries occur.
- The architecture determines the choice of genetic code, assembler, mapping strategy, gene predictor, treatment of repeats and validation data.
- An assembly or annotation is a model of a sampled genome, not the genome itself.

Common errors:

- Stating that all bacteria have one circular chromosome.
- Treating all eukaryotic transcripts as monocistronic, or all viral genomes as unspliced.
- Taking a circular contig as proof that the molecule is a single circle in the cell.
- Using the Baltimore classification as if it were a taxonomy of viruses.
- Equating compactness with simplicity.

Connections: Section 6.4; assembly in Section 5.3; gene prediction in Section 5.6; phylogenetic classification in Section 4.5.

6.14.2 Introns in bacteria and archaea

Correct the statement that prokaryotes lack introns, and explain how the corrected picture changes sequence analysis and annotation.

Extensions, solution outline and common errors

Extensions:

1. Compare spliceosomal introns with self-splicing group I and group II introns.
2. Why can similarity of primary sequence fail to detect self-splicing introns?
3. How would you distinguish an intron from an insertion error in an assembly?
4. How can introns disrupt the annotation of protein-coding and RNA genes?

Solution outline:

- Spliceosomal introns are a feature of eukaryotic nuclear pre-mRNAs. Introns in bacteria and archaea are rare and of other kinds: self-splicing group I and group II introns, and archaeal introns in tRNA and rRNA genes that are removed by a dedicated enzyme. Group II introns are thought to be the ancestors of spliceosomal introns.
- The RNA secondary structure and catalytic core are more conserved than the primary sequence, so covariance models, which score base-pair covariation as well as sequence, detect them better than sequence similarity alone.
- An intron should be supported by spliced transcripts, an alignment of the flanking exons with the host gene in related species, a plausible RNA structure and splice boundaries, and consistent read support, after excluding misassembly and contamination.
- An intron can split an open reading frame, obscure an rRNA or tRNA gene, or be mistaken for an inserted mobile element.
- A low frequency does not mean absence; classify each intron only as far as the evidence allows.

Common errors:

- Replacing one universal claim with another, that all prokaryotes contain self-splicing introns.
- Calling all introns spliceosomal.
- Treating sequence similarity alone as sufficient evidence of homology for structured RNAs.
- Ignoring the homing endonucleases encoded by many self-splicing introns, which make them mobile.
- Inferring an intron from an unexplained gap in a protein alignment.

Connections: Section 6.4; RNA structure and covariance models in Section 1.7; gene prediction in Section 5.6.

6.14.3 Interpreting regulatory and chromatin evidence

A distal region is accessible, enriched for the histone mark H3K27ac, bound by a transcription factor, and in contact with a promoter in Hi-C data. Which conclusions are justified, and which experiments would show that it is an enhancer of that gene?

Extensions, solution outline and common errors

Extensions:

1. What does each assay measure, and what does it not measure?
2. Why is contact frequency not the same as regulation?
3. What is a candidate cis-regulatory element?
4. Distinguish tests of necessity, sufficiency and target assignment.
5. How do cell type, developmental stage and mixtures of cells affect the interpretation?

Solution outline:

- Interpret each signal separately: accessibility shows open chromatin; H3K27ac is a mark associated with active elements; factor binding is measured in the conditions assayed; contact frequency shows that the regions are close more often than expected in a population of cells.
- None of these alone shows that the region regulates the promoter. The signals can reflect a shared cell state, or be a consequence of transcription rather than its cause.

- *Candidate* keeps the distinction between a biochemical annotation and a demonstrated function. State the cell type, genome build, processing pipeline and thresholds.
- Test necessity at the endogenous locus by deletion, CRISPR interference or mutation of the binding motif; test sufficiency by targeted activation or reporter assays, remembering that an ectopic context can change activity.
- Assign the target gene through the response to perturbation, allele-specific effects and genetic evidence, with positive and negative controls and alternative candidate promoters.
- Report a graded conclusion: biochemically active candidate, putative target, perturbation-supported regulator, or mechanistically resolved enhancer.

Common errors:

- Calling every accessible region an enhancer.
- Treating histone modifications as definitive functional labels.
- Assigning the nearest gene as the target without further evidence.
- Treating reporter activity as proof of function at the endogenous locus.
- Pooling tissues or cells without accounting for the mixture.

Connections: Section 6.9; 3D genome organisation in Section 6.7; expression data in Chapter 7.

6.14.4 Biological repeats and repeat-induced artefacts

Explain why repeated sequence is both a major component of genome biology and a source of errors in assembly, mapping and annotation. Outline a strategy for repeat-aware gene annotation.

Extensions, solution outline and common errors

Extensions:

1. Distinguish tandem repeats, interspersed repeats, segmental duplications and transposable elements.
2. What are autonomous and non-autonomous transposable elements?
3. How do repeats affect short-read mapping and genome assembly?
4. Compare hard and soft masking.
5. How would you avoid annotating a transposable element as a host gene?

Solution outline:

- Define repeat classes by arrangement, homology and mechanism. Repeated sequence need not be mobile, and a duplicated host gene is not a transposable element.
- Autonomous elements encode the proteins needed for their own movement, such as L1; non-autonomous elements, such as Alu, depend on proteins encoded elsewhere. Most copies in a genome are inactive, decayed or nested, and some have been domesticated.
- Similar copies cause reads to map to several places, assemblies to collapse or fragment, and false duplications and ambiguous variant calls. Long accurate reads help but do not resolve every array.
- Hard masking removes repeats from many analyses, improving specificity at the cost of missed genes; soft masking keeps the sequence but lets gene predictors down-weight repeat-derived candidates.
- Combine a curated repeat library for the lineage with de novo discovery of repeat families, and examine terminal repeats, transposase or reverse-transcriptase domains, copy number, splicing, conservation and protein evidence.

- Keep host genes that overlap repeats only when there is evidence for a distinct transcript and coding structure; transcription alone does not show function.

Common errors:

- Treating repeats as uniformly non-functional.
- Interpreting transcription as proof of a host-gene function.
- Assuming that every copy of a transposable element is active.
- Treating a repeat library from a distant species as complete.
- Ignoring genes that overlap or evolved from transposable elements.

Connections: Section 6.7; assembly in Section 5.3; mapping in Section 5.4; gene annotation in Section 5.6.

6.14.5 Building and interpreting a bacterial pangenome

You build a pangenome from 500 bacterial isolates. Explain how sampling, assembly quality, the clustering of genes into families and population structure affect its interpretation.

Extensions, solution outline and common errors

Extensions:

1. Define core and accessory gene families operationally.
2. How do annotation errors inflate the accessory genome?
3. Compare a table of gene content with a sequence graph.
4. Evaluate a claim that the pangenome is open or closed.
5. Design an analysis of association between accessory genes and virulence.

Solution outline:

- A gene-content pangenome is the union of gene families in a stated sample; a sequence pangenome represents alternative sequence paths. They answer different questions.
- Use consistent assembly and annotation, then cluster proteins with rules that consider coverage, domains and orthology. Build presence–absence and copy-number tables, and inspect difficult families with trees and conserved gene order.
- For n genomes, the prevalence of family g is $p_g = \frac{1}{n}\sum_i \mathbf{1}(A_{gi} > 0)$, where A_{gi} is the number of copies in genome i; state the threshold used for core genes. Missing sequence, fragmented genes and split or merged clusters all change p_g.
- Estimate the uncertainty of the accumulation curve by repeated subsampling or permutation of the genome order. Whether a pangenome is open or closed depends on the sample, the fitted model and the technical errors.
- Use a graph when haplotypes, genomic context or mapping are central, and gene tables for questions about family presence and copy number. Neither removes sampling bias.
- For an association with virulence, model population structure, relatedness, recombination, environment, missing data and multiple testing, and follow the association with genetic and phenotypic validation.

Common errors:

- Treating a 95% prevalence threshold as universal.

- Clustering proteins by percentage identity alone.
- Interpreting every singleton as a new gene.
- Assuming that a graph is unbiased because it is not linear.
- Ignoring clonal population structure in association studies.

Connections: Section 6.6; horizontal gene transfer in Section 3.7; pangenome graphs and mapping in Section 5.4; multiple testing in Section 7.4.1.

6.14.6 Interpreting conserved synteny

Two genes lie in corresponding genomic neighbourhoods in human and in a fish. Which evolutionary inferences does this observation support?

Extensions, solution outline and common errors

Extensions:

1. Is synteny the same as identical gene order?
2. Can orthologues lack conserved synteny?
3. How does synteny reveal a whole-genome duplication?
4. How would you score a syntenic block?
5. Which technical artefact can mimic a breakpoint?

Solution outline:

- Define the terms: synteny may mean correspondence of chromosomes or regions, whereas collinearity also requires conserved order and orientation.
- Corresponding neighbours support shared positional ancestry and help to identify the orthologue among paralogues and to find missed genes, but rearrangements can move true orthologues.
- Choose orthologous anchors, chain them into blocks with stated rules for orientation and gaps, and compare the result with randomised gene orders. Use chromosome-scale assemblies and outgroups.
- An ancient whole-genome duplication leaves pairs of paralogous segments with complementary retained genes, even after most duplicates have been lost; teleost fish had an additional duplication, so one human region often corresponds to two fish regions. Conserved synteny has helped to resolve the early vertebrate duplications (Simakov et al. 2020).
- Confirm breakpoints with long reads, maps or independent assemblies, because misjoined scaffolds and gaps mimic rearrangements.

Common errors:

- Using “synteny” without an operational definition.
- Treating one neighbouring gene as a convincing block.
- Requiring conserved order for all orthologues.
- Ignoring tandem duplicates when chaining anchors.
- Comparing coordinates from different assembly releases.

Connections: Section 6.5; genome duplications in Section 6.4 and Section 3.4; orthology databases in Section 3.4; long-read assembly in Section 5.3.

6.14.7 A short ORF without homologues

Ribosome profiling of human heart tissue shows translation of a 40-codon ORF in a transcript annotated as a long non-coding RNA. A BLAST search finds no homologues outside primates. Is this a new protein-coding gene, and is it a *de novo* gene?

Extensions, solution outline and common errors

Extensions:

1. Why was the ORF not in the protein-coding annotation?
2. What does the Ribo-seq signal show, and what does it not show?
3. How would you show that the protein exists?
4. How would you distinguish *de novo* origin from homology detection failure?
5. What can a predicted structure contribute?

Solution outline:

- Classical annotation used minimum ORF lengths, often 100 codons, because short ORFs are common in random sequence, and short ORFs give weak gene-finder and conservation signals.
- Check the Ribo-seq evidence: three-nucleotide periodicity in frame, footprints across the whole ORF, a defined start, and reproducibility across samples. Translation shows that ribosomes engage the ORF, not that a stable, functional protein is made; it may be a regulatory uORF-like event or noise.
- Look for the protein with mass spectrometry (targeted or peptidomic searches, immunopeptidomics), with strict, separate error control for non-canonical peptides, or with tagging of the endogenous locus.
- For function, use knockouts or knockdowns that disrupt the ORF but not the RNA (for example a start-codon mutation), and evolutionary signals such as PhyloCSF or dN/dS within primates.
- For origin, find the orthologous region in other mammals by conserved gene order and whole-genome alignment, check whether the ORF is absent (disrupted or untranscribed) there, and identify the enabling mutations. Absence of BLAST hits alone is not evidence of *de novo* origin.
- A structure prediction with low confidence is expected for a short, young protein; a confident complex prediction with a plausible partner is a hypothesis to test experimentally.

Common errors:

- Treating ribosome footprints as proof of a functional protein.
- Treating absence of BLAST hits as proof of a *de novo* origin.
- Accepting a single non-canonical peptide at the global FDR.
- Interpreting a confident structure model of a short peptide as evidence that it folds or functions.

Connections: Section 6.8; gene prediction in Section 5.6; proteogenomics and error control in Section 8.3; structure prediction in Section 9.8; homology detection in 1.

6.14.8 Inferring how a deletion arose

Long-read sequencing of two unrelated patients with the same neurological disorder shows a deletion in each. In one, the deletion is 1.4 Mb and its ends lie in two nearly identical 24 kb repeats. In the other, the deletion is 600 kb, its junction shows a 3 bp microhomology, and a 150 bp segment copied from a nearby region is inserted at the junction. Which mechanisms are likely, and what do they imply?

Extensions, solution outline and common errors

Extensions:

1. Which mechanism produces recurrent rearrangements, and why?
2. Why is the exact junction hard to determine in the first patient?
3. Which gene could explain the shared phenotype?
4. How would you decide whether a new deletion in a third patient is likely to be pathogenic?

Solution outline:

- The first deletion, flanked by segmental duplications in direct orientation, is typical of NAHR. NAHR events recur with the same breakpoints in unrelated individuals and produce reciprocal duplications.
- Within the long repeats, the crossover point can only be placed in the stretch where the two copies differ; short reads cannot map it at all, and even long reads must span the repeat.
- The second deletion is non-recurrent. Microhomology alone suggests NHEJ or microhomology-mediated repair; a templated insertion from nearby suggests a replication-based mechanism (FoSTeS/MMBIR).
- Both deletions remove a common region; a dosage-sensitive (haploinsufficient) gene in the overlap is the likely cause. Check constraint and dosage annotations for the genes in the region.
- For a new deletion, compare with population SV catalogues (gnomAD-SV), dosage-sensitivity databases and segregation in the family, and confirm the breakpoints.

Common errors:

- Assuming that all recurrent deletions are caused by NHEJ.
- Treating the deletion size as evidence of the mechanism.
- Assuming that every gene in a deleted region contributes to the phenotype.

Connections: Section 6.7; SV detection in Section 5.4; dosage balance in Section 3.4.

7 Expression analysis

> **From the exam description**
>
> Expression analysis of genes and proteins: methods, algorithms, statistics.

7.1 Overview

Measuring which genes are expressed, and how much, links the genome to the phenotype. Expression analysis was the first area of genomics that produced large, quantitative data sets about many genes in many samples, and it forced biologists to confront statistical problems that are now everywhere: normalisation, multiple testing, batch effects and the exploration of high-dimensional data. This chapter follows the development from microarrays to single-cell and spatial methods:

- microarrays, their preprocessing and what they taught us (Section 7.2);
- RNA-seq: library preparation and experimental design, spliced alignment, transcript assembly, quantification and normalisation (Section 7.3);
- statistics: differential expression with count models, multiple testing, clustering and its validation, classification with many genes and few samples, and dimensionality reduction (Section 7.4);
- interpretation: gene set analysis, co-expression and eQTLs, and the relation between mRNA and protein levels (Section 7.5);
- single-cell genomics: technologies, analysis, multimodal data, spatial transcriptomics, foundation models and cell types (Section 7.6);
- single-cell sequencing of cancer and tumour evolution (Section 7.7).

The databases that archive expression data are described in Section 2.8, and the measurement of proteins in Chapter 8.

7.2 The first genome-wide methods: microarrays

Measuring which genes are expressed, and how much, links the genome to the phenotype. Before the 1990s, expression was measured one gene at a time, for example with Northern blots. Two approaches made it possible to measure all genes at once: counting short sequence tags, as in *serial analysis of gene expression* (SAGE) (Velculescu et al. 1995), and hybridisation to DNA microarrays. Microarrays dominated for about fifteen years, and they introduced most of the ideas that are still used to analyse expression data.

7.2.1 Principle

A **DNA microarray** is a solid surface with thousands to millions of spots (*features*), each containing many copies of a DNA probe for one transcript. RNA from a sample is converted to labelled cDNA or cRNA and hybridised to the array. Each labelled molecule binds to the probe with the complementary

sequence, and the fluorescence intensity of each spot is a measure of how abundant that transcript was in the sample. The relation is monotonic but not simply proportional: background dominates at low abundance, the signal saturates at high abundance, and each probe has its own binding affinity. Intensities are therefore relative measures, best compared for the same probe between samples. Three main technologies were used:

- **Spotted cDNA arrays**, developed in Patrick Brown's laboratory at Stanford, were made by robotically printing PCR products onto glass slides (Schena et al. 1995). They were usually used in *two-colour* experiments: two samples, labelled with different dyes (Cy3 and Cy5), were hybridised to the same array, and the result was the ratio between them. Because they could be built in academic laboratories, they spread rapidly.
- **Affymetrix GeneChips** were made by synthesising short (25-nucleotide) oligonucleotides directly on the chip with photolithography, the technique used to make computer chips (Fodor et al. 1991; Lockhart et al. 1996). Each gene was measured with a *probe set* of about 11–20 different probes, each with a *mismatch* control probe that differed in the central base. These were *one-colour* arrays, each measuring one sample, and the high reproducibility of the industrial process made them the most widely used platform.
- **Illumina BeadArrays** use a different principle. Probes are attached to microscopic beads, which are randomly distributed in wells etched in a fibre-optic bundle or a glass slide. Since the position of each bead is random, the array is decoded after manufacture by a series of hybridisations that identify which probe each bead carries (Gunderson et al. 2004). Each probe is present on many beads, which gives built-in replication. BeadArrays were used for expression profiling (Kuhn et al. 2004), and, above all, for **SNP genotyping**: the genotyping chips used in most genome-wide association studies, measuring up to millions of SNPs per sample, are Illumina bead arrays.

7.2.2 Preprocessing and normalisation

The raw data from an array are images, and many steps are needed before intensities can be compared between genes and samples (Quackenbush 2001):

1. **Image analysis**: locating the spots and estimating their intensity and the local background.
2. **Background correction**: removing the signal from non-specific hybridisation and the optical background.
3. **Transformation**: working on the logarithmic scale, usually $\log_2$, because intensities span several orders of magnitude, errors are roughly multiplicative, and fold changes become differences.
4. **Normalisation**: removing systematic differences between arrays or dyes, caused by differences in the amount of RNA, labelling efficiency or scanner settings. For two-colour arrays, the log-ratio $M = \log_2(R/G)$ is plotted against the average log intensity A (an *MA plot*), and a curve fitted with local regression (loess) is subtracted, so that most genes have M close to zero. For one-colour arrays, **quantile normalisation** forces all arrays to have the same distribution of intensities (Bolstad et al. 2003).
5. **Summarisation**: for Affymetrix arrays, combining the probes of a probe set into one expression value. The *robust multi-array average* (RMA) fits a model across arrays that accounts for probe-specific affinities, and ignores the mismatch probes, which turned out to measure specific signal as well as background (Irizarry 2003).

These steps are not just technicalities. Normalisation methods assume that most genes do not change between samples, an assumption that fails in some experiments, and the choice of method can change the results.

7.2.3 What microarrays showed, and their limitations

Microarrays made it possible to ask questions about whole transcriptomes. Clustering genes by their expression profiles showed that genes with related functions are co-expressed (Eisen et al. 1998). Expression profiles could classify cancers: acute leukaemias could be told apart by expression alone (Golub et al. 1999), breast tumours fell into molecular subtypes with different prognoses (Perou et al. 2000), and a 70-gene signature predicted the outcome of breast cancer (Veer et al. 2002) and became a clinical test. Atlases of expression across tissues showed which genes are ubiquitous and which are tissue specific (Su et al. 2004). The analysis of these experiments raised the statistical problems that are discussed in Section 7.4: how to test thousands of genes with few replicates, for which empirical Bayes methods such as *limma* were developed (Smyth 2004), and how to control false discoveries.

The early years also showed how easily such experiments go wrong. Results were hard to reproduce between laboratories and platforms until standards for reporting experiments, MIAME (Brazma et al. 2001), and large comparison studies (MAQC Consortium 2006) were established. Microarrays have several built-in limitations:

- they can only measure transcripts whose sequences are known and represented on the array;
- signals are affected by *cross-hybridisation* of similar sequences and by probe-specific affinities, so intensities of different genes cannot be compared directly;
- the dynamic range is limited by background at the low end and saturation at the high end;
- they cannot easily distinguish isoforms, detect new transcripts or measure allele-specific expression.

RNA-seq (Section 7.3) removed most of these limitations, and replaced microarrays for expression profiling from about 2010. Arrays remain in use for genotyping and DNA methylation profiling, where they are cheap and robust.

See also: further explanation in Section D.7.1; related Wikipedia articles in Section E.7.1.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles DNA microarray, Gene expression profiling, Serial analysis of gene expression, Quantile normalization and MA plot (CC BY-SA 4.0, retrieved 4 October 2026).

7.3 RNA sequencing

RNA-seq measures expression by sequencing cDNA made from the RNA in a sample and counting the reads that come from each gene or transcript. It was introduced in 2008 (Mortazavi et al. 2008; Nagalakshmi et al. 2008), shown to agree well with microarrays while being more reproducible and sensitive (Marioni et al. 2008), and quickly became the standard (Wang et al. 2009; Stark et al. 2019). Because it does not depend on predefined probes, RNA-seq can discover new genes and transcripts, measure isoforms, detect allele-specific expression, fusion genes and RNA editing, and be used in any organism, with or without a reference genome. Its dynamic range is limited only by the sequencing depth.

7.3.1 Library preparation and experimental design

Most of the RNA in a cell is ribosomal. To sequence mRNA, either polyadenylated RNA is selected with oligo(dT), or rRNA is depleted, which also keeps non-polyadenylated RNAs. The RNA is fragmented,

reverse transcribed into cDNA, ligated to adapters and amplified. *Strand-specific* protocols record which strand the transcript came from, which matters where genes overlap on opposite strands. Typical experiments use 20–50 million short reads per sample for gene-level analysis. Long-read sequencing (Section 5.2) reads entire transcripts and is increasingly used to define isoforms (see below).

The most important design decision is the number of **biological replicates**. The key concept is the **experimental unit**: the smallest unit that is independently assigned to (or sampled from) a condition, such as an animal, a patient or an independently grown culture. *Biological replicates* are different experimental units in the same condition, and they capture the natural variation between individuals. *Technical replicates* are repeated measurements of the same unit, for example the same RNA sequenced in two libraries or two lanes; they measure only the noise of the protocol. Technical replicates show little variation, while biological replicates vary much more, and differential expression can only be assessed relative to biological variation. Treating technical replicates, or several samples from the same individual, as independent biological replicates is called *pseudoreplication* and gives far too many significant results. Studies have shown that with three biological replicates many true changes are missed, and that six or more are needed for good sensitivity (Schurch et al. 2016).

Good design also controls **confounding**. Technical *batch effects* (different days, operators, reagent lots or sequencing runs) can be larger than the biological differences of interest (Leek et al. 2010). If all control samples are processed in one batch and all treated samples in another, batch and condition are confounded, and no statistical method can separate them. The remedies are classical: **randomise** the assignment of samples to conditions and to processing order; **block**, by processing equal numbers of samples from each condition in every batch, so that batch can be included as a covariate in the model (Section 7.4); and use **paired designs** where possible, for example tumour and normal tissue from the same patient, or before and after treatment in the same individual, so that the comparison is made within each individual and differences between individuals cancel out. Such design decisions must be made before the experiment; they cannot be repaired in the analysis.

7.3.2 The analysis workflow

A standard analysis has the following steps (Conesa et al. 2016):

1. **Quality control** of the reads: base qualities, adapter content, duplication and rRNA contamination.
2. **Alignment or quantification**, described below.
3. **Counting** reads per gene, or estimating abundances per transcript.
4. **Normalisation** for sequencing depth and composition.
5. **Exploratory analysis**, such as PCA and clustering, to detect outliers and batch effects (Section 7.4).
6. **Statistical testing** for differential expression, and interpretation with gene sets and pathways (Section 7.5).

Each step rests on assumptions, and an observed difference can have several explanations besides regulation within cells: a change in cell composition, poor RNA quality, protocol bias, errors in the annotation or a batch effect (Figure 7.1). A longer gene list does not repair a flawed design.

7.3.2.1 Quality control

Quality control is a diagnosis of how the data were generated, not an automatic filter. Before alignment, base qualities, adapter content, over-represented sequences and per-cycle base composition are checked (Section 5.4). After alignment, the useful statistics are specific to RNA: the fraction of uniquely mapped reads, the fractions of ribosomal and mitochondrial reads, strand specificity, splice-junction support and the coverage along gene bodies. Degraded RNA gives a characteristic **3′ bias**, since fragments

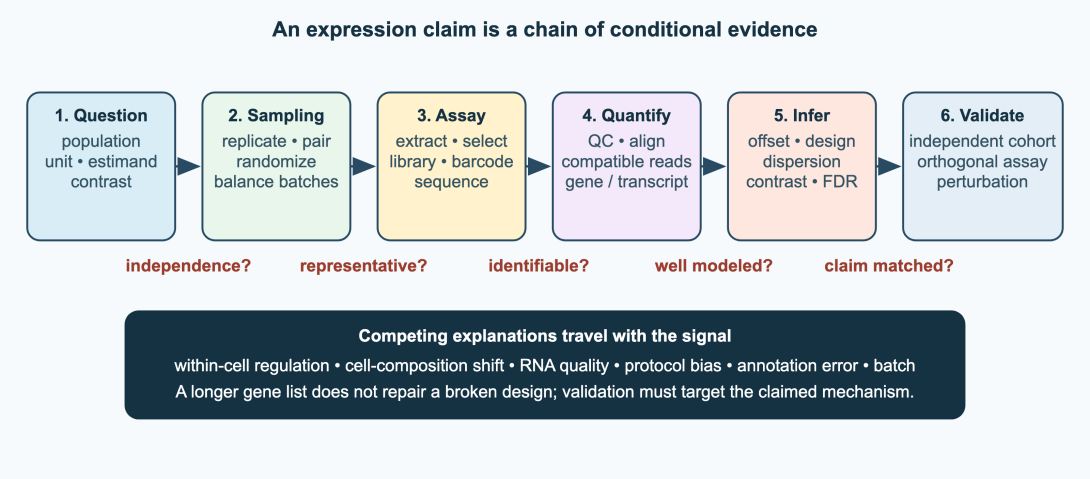


Figure 7.1: An expression result as a chain of evidence, from the biological question to validation. Each step has its own question that must be answered before the next can be trusted. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

near the poly(A) tail survive best. At the level of samples, PCA, sample correlations and clustering show whether samples group by condition or by processing batch, and genotypes called from the reads can confirm that each sample comes from the expected individual. An outlying sample is a reason to investigate, not to exclude: exclusion rules should be decided before the results are known, and conclusions checked with and without doubtful samples. The SEQC/MAQC-III consortium sequenced reference RNA mixtures at many sites and showed that platform, site, expression level and analysis pipeline all affect the results, while relative expression measures were reproducible across sites ("A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium" 2014).

7.3.2.2 Spliced alignment

RNA-seq reads come from spliced transcripts, so a read that spans an exon–exon junction aligns to the genome in two parts separated by an intron, which may be tens of kilobases long. Ordinary DNA read mappers (Section 5.4) cannot handle this. **TopHat** first mapped reads that align without gaps, predicted junctions from them and then aligned the remaining reads across those junctions (Trapnell et al. 2009). **STAR** searches for the longest prefix of the read that matches the genome exactly (the *maximal mappable prefix*) using an uncompressed suffix array, then repeats the search for the rest of the read, and stitches the pieces together (Dobin et al. 2012). It is very fast but needs about 30 GB of memory for the human genome. **HISAT2** uses a hierarchical graph FM-index with much smaller memory requirements (Kim et al. 2019).

7.3.2.3 Transcript assembly

Alignments can be used to reconstruct the transcripts themselves. **Reference-guided assembly**, with Cufflinks (Trapnell et al. 2010) or StringTie (Pertea et al. 2015), builds a graph of overlapping alignments for each gene and finds a small set of transcripts that explain the reads, which can reveal new isoforms. Cufflinks used a minimum path cover of the overlap graph, the most parsimonious set of transcripts. In organisms without a reference genome, transcripts are assembled **de novo** with de Bruijn graph methods such as Trinity (Grabherr et al. 2011) (Section 5.3).

7.3.2.4 Quantification and pseudo-alignment

The simplest quantification counts the reads that overlap each gene. Estimating the abundance of individual **transcripts** is harder, because isoforms of the same gene share most of their exons, so many reads are compatible with several transcripts. RSEM solves this with the **expectation–maximisation** (EM) algorithm: it alternates between assigning each read fractionally to transcripts in proportion to their current abundance estimates, and re-estimating the abundances from these assignments (Li and Dewey 2011).

Aligning every read base by base is not necessary for quantification. It is enough to know *which* transcripts a read is compatible with. **kallisto** introduced **pseudo-alignment**: it builds a de Bruijn graph of the transcriptome, looks up the k-mers of each read, and intersects the sets of transcripts that contain them (N. L. Bray et al. 2016). Reads with the same set of compatible transcripts form an *equivalence class*, and the EM algorithm then runs on the counts of these classes. **Salmon** uses a similar *selective alignment* and models sequence-specific, fragment-GC and positional biases (Patro et al. 2017). These methods are one to two orders of magnitude faster than alignment and similarly accurate for known transcripts, but they cannot discover new ones.

7.3.2.5 Long-read RNA-seq

Short reads cover only a small part of a transcript, so the full isoform must be inferred by assembly or by EM, and different combinations of exons far apart in the same gene are hard to resolve. Long-read platforms (PacBio and Oxford Nanopore; Section 5.2) can sequence complete cDNA molecules, and nanopore sequencing can also read native RNA directly. Each read then shows one **full-length isoform**, with its exact combination of exons, its start and end sites and its poly(A) tail, so no assembly is needed. Reads are aligned to the genome with spliced aligners such as minimap2 (Li 2018), and grouped into known and new transcripts. This has made long reads the method of choice for **transcript discovery** and for building isoform catalogues, and has shown that many genes have more isoforms than annotated (Pardo-Palacios et al. 2024). The drawbacks are fewer reads per sample, higher error rates on some platforms (notably direct RNA sequencing), which make exact splice sites harder to call, truncated reads from degraded RNA, and a bias towards shorter molecules. Short reads are therefore still often used for quantification, sometimes in combination with long reads.

7.3.3 Normalisation

Raw counts are not comparable between genes or samples. Longer transcripts produce more fragments, and samples sequenced more deeply produce more reads. A gene with twice as many reads as another is therefore not necessarily twice as highly expressed. Several units correct for these effects:

- **CPM** (counts per million) divides by the library size and corrects for depth only.
- **RPKM/FPKM** (reads or fragments per kilobase of transcript per million mapped reads) also divides by the transcript length (Mortazavi et al. 2008).
- **TPM** (transcripts per million) first divides each count by the transcript length, and then scales so that the values in each sample sum to one million:

$$\mathrm{TPM}_i = \frac{c_i/\ell_i}{\sum_j c_j/\ell_j} \times 10^6,$$

  where c_i is the count and ℓ_i the effective length of transcript i. TPM estimates the relative molar concentration of each transcript and, unlike RPKM, sums to the same total in every sample (Wagner et al. 2012).

These units do not handle **composition bias**. RNA-seq data are *compositional*: a library contains a fixed number of reads sampled from the RNA in it, so the counts measure each gene's share of the library, not its absolute amount. If a few genes are very highly expressed in one sample, they take up a large share of the reads, and all other genes appear lower even if their true expression is unchanged. Methods for differential expression therefore estimate normalisation factors under the assumption that most genes are not differentially expressed:

- **TMM** (trimmed mean of M-values), used by edgeR, computes the weighted mean of log-ratios between a sample and a reference sample after trimming the most extreme genes (Robinson and Oshlack 2010).
- The **median-of-ratios** method of DESeq computes, for each gene, the ratio of its count to the geometric mean of its counts across all samples. The median of these ratios in a sample is its *size factor* (Anders and Huber 2010; Love et al. 2014).

This assumption can fail. If the total amount of mRNA per cell changes, for example when a transcription factor such as MYC raises the expression of most genes, every gene's share of the library stays about the same, and all of these normalisations wrongly conclude that little has changed (Lovén et al. 2012). Detecting such global shifts requires an external reference, such as synthetic *spike-in* RNAs added in proportion to the number of cells.

TPM values are useful for comparing the relative abundance of transcripts within a sample. Because TPMs sum to the same total in every sample, they are still proportions, and they do not remove composition effects between samples; they are suitable for descriptive comparisons but are not a general normalisation. For differential expression, tools such as DESeq2 and edgeR take the raw (unnormalised) integer counts as input, not TPM or CPM values, and include the normalisation factors as offsets in the statistical model (Section 7.4).

See also: further explanation in Section D.7.2; related Wikipedia articles in Section E.7.2.

Sources

Parts of this section are adapted from the Wikipedia articles RNA-Seq, Transcriptomics technologies and Expectation–maximization algorithm (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Gene-Expression Analysis" (CC BY 4.0), used with permission.

7.4 Testing and statistics

Expression studies measure tens of thousands of genes in a handful of samples. This raises three statistical problems that appear throughout genomics: modelling the variability of the measurements with few replicates, testing thousands of hypotheses at once, and exploring high-dimensional data. The section starts with the general statistical background that is expected of every student, whatever their topic.

7.4.1 Statistical background

Most bioinformatics results end with a statistical statement: an E-value, a p-value for differential expression, a confidence in a phylogenetic split, an accuracy on a test set. You should be comfortable with the basic concepts and know their common pitfalls. The Nature Methods *Points of Significance* series (Statistics for Biologists) covers all of this in two-page articles and is the recommended starting point (Krzywinski and Altman 2013).

7.4.1.1 Probability and distributions

You should be able to explain conditional probability and **Bayes' theorem**,

$$P(H \mid D) = \frac{P(D \mid H)\,P(H)}{P(D)},$$

and use it on a simple example. A classic one: a test that detects a rare condition (1 in 1,000) with 99% sensitivity and 99% specificity still gives mostly false positives, because the few true cases are outnumbered by the 1% false positives among the many healthy people. The same base-rate argument explains why E-values must account for database size (Section 1.4) and why most predicted interactions or binding sites in a genome-wide screen are false.

Know the common distributions and when they arise: the **binomial** (number of successes in n trials), the **Poisson** (rare events, read counts at low coverage), the **normal** (sums of many small effects), the **negative binomial** (overdispersed counts, as in RNA-seq; Chapter 7) and the **extreme value** distribution (maxima, as in alignment scores; Section 1.4). Understand **likelihood**, the probability of the data given a model, and maximum-likelihood estimation. It underlies phylogenetics (Chapter 4), HMM training (Section 1.6) and much of machine learning.

7.4.1.2 Hypothesis tests and p-values

A statistical test compares the data with a **null hypothesis**. The **p-value** is the probability, *assuming the null hypothesis is true*, of observing data at least as extreme as those seen. It is not the probability that the null hypothesis is true, and a large p-value does not show that the null is true. A small p-value says nothing about how large or important an effect is: with enough data, a trivial difference becomes "significant". Report **effect sizes** with **confidence intervals**, not only p-values (Krzywinski and Altman 2013; Nuzzo 2014; Wasserstein and Lazar 2016).

You should know the most common tests and their assumptions:

- the **t-test** compares means. The one-sample test compares a mean with a fixed value, the two-sample test compares two groups, and the paired test compares two measurements on the same units. It assumes approximately normal data (or large samples) and, in its classic form, equal variances; **Welch's t-test** drops the equal-variance assumption and is usually the better default. The test statistic is the difference in means divided by its standard error;
- **non-parametric** alternatives, such as the Mann–Whitney (Wilcoxon rank-sum) and Wilcoxon signed-rank tests, use ranks and make fewer assumptions, at some cost in power;
- **χ^2 and Fisher's exact tests** for counts in contingency tables, for example enrichment of a GO term in a gene list (Chapter 10);
- **correlation** (Pearson for linear relationships, Spearman for monotonic ones), and why correlation does not imply causation;
- **permutation tests**, which build the null distribution by shuffling labels, and are useful when no standard test fits.

Know also the difference between **type I** errors (false positives, controlled by the significance level α) and **type II** errors (false negatives), and what determines **statistical power**: effect size, variability and sample size. Too few replicates is the most common reason for unreliable conclusions in genomics experiments.

7.4.2 Differential expression

The basic question is which genes change in expression between conditions. A t-test for each gene works poorly. With three replicates per group, the variance of each gene is estimated very imprecisely, and genes that happen to have a small sample variance get spuriously large t-statistics.

Moderated statistics. The solution, developed for microarrays, is to borrow information across genes. In the *limma* framework (Smyth 2004), an empirical Bayes model treats the true variances of all genes as drawn from a common prior distribution, estimated from all genes together. Each gene's variance is then *shrunk* towards the common value. The resulting moderated t-statistic is much more stable, and limma remains in use, also for RNA-seq data after a transformation that estimates the mean–variance relationship (voom) (Law et al. 2014).

Count models. RNA-seq produces counts. If the same library were sequenced repeatedly, the count of a gene would follow a Poisson distribution, with variance equal to the mean (Marioni et al. 2008). Between biological replicates, the variance is larger (*overdispersion*). The standard model is the **negative binomial** distribution, with variance

$$\mathrm{Var}(K) = \mu + \alpha\mu^2,$$

where μ is the mean and α the gene-specific *dispersion*. For highly expressed genes, the second term dominates, and $\sqrt{\alpha}$ is the biological coefficient of variation. edgeR (Robinson et al. 2009) and DESeq2 (Love et al. 2014) fit a **generalised linear model** for each gene, with a log link,

$$\log \mu_{ij} = \log s_j + \sum_r x_{jr}\beta_{ir},$$

where s_j is the normalisation factor of sample j (Section 7.3), x_{jr} the design matrix describing the samples (condition, batch, paired individuals) and β_{ir} the coefficients, such as the log fold change between conditions. As in limma, the dispersion of each gene is estimated with shrinkage towards a trend fitted across all genes. DESeq2 also shrinks the *fold changes*, so that large fold changes of low-count, noisy genes are not over-interpreted. Significance is assessed with Wald or likelihood-ratio tests. Including known batches and other covariates in the design matrix is the main way to deal with batch effects (Leek et al. 2010).

The tests answer slightly different questions. A **Wald test** divides an estimated coefficient, or a combination of coefficients, by its standard error. A **likelihood-ratio test** compares a full model with a reduced model without the terms of interest, and is the natural choice when the question involves several coefficients at once, such as any difference between several time points. The **quasi-likelihood F-test** of edgeR also takes into account the uncertainty in the estimated dispersion, and gives better control of false positives when there are few replicates (Lund et al. 2012). The test should follow from the question, not be chosen because it gives the most significant genes.

7.4.3 Design matrices, contrasts and interactions

The **design matrix** has one row per sample and one column per coefficient, and it states which comparisons the model can make. In a **paired design**, such as treated and untreated samples from the same donors, the model `~ donor + treatment` includes one coefficient per donor, so that the treatment effect is estimated *within* donors, and stable differences between donors are removed. A known batch is added in the same way, as long as it is not confounded with the condition. If two columns of the design matrix are identical, for example when all treated samples were processed in one batch,

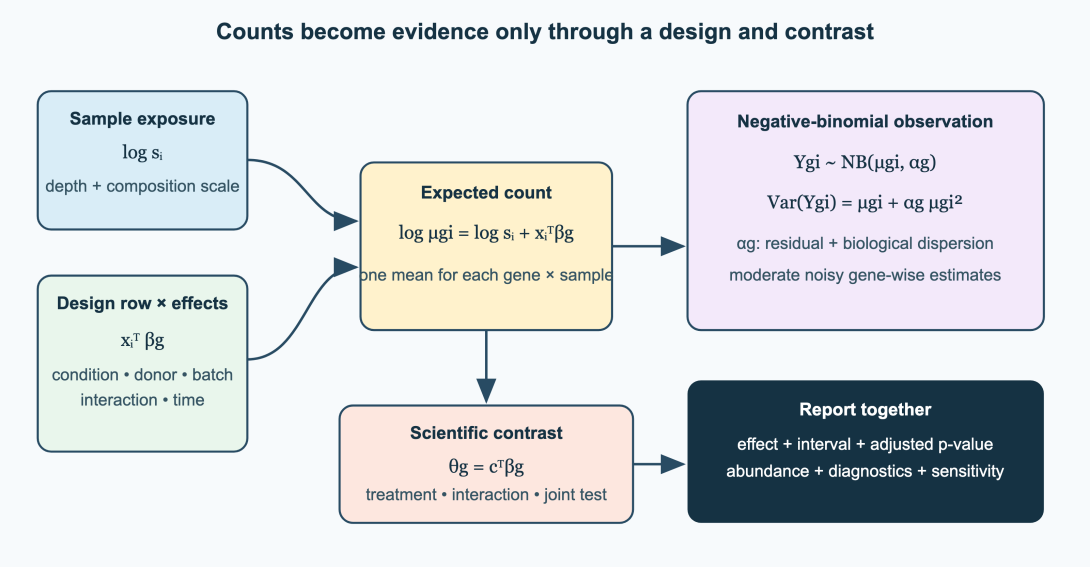


Figure 7.2: The negative binomial model of RNA-seq counts. The expected count of each gene in each sample combines a sample-specific offset for depth and composition with the design matrix and coefficients; the observed count varies around it with a gene-specific dispersion; a contrast of the coefficients answers the biological question. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

the matrix is not of full rank and the two effects cannot be separated. Software that refuses to fit such a model is doing the right thing.

A **contrast** is a linear combination of coefficients, $c^T\beta$, that answers one question. In an experiment with two genotypes, A and B, each with and without treatment, the model `~ genotype + treatment + genotype:treatment` includes an **interaction** term. The interaction is a difference of differences:

$$(\mathrm{B}_{\mathrm{treated}} - \mathrm{B}_{\mathrm{control}}) - (\mathrm{A}_{\mathrm{treated}} - \mathrm{A}_{\mathrm{control}}),$$

and asks whether the response to treatment depends on the genotype. It is not the treatment effect in either genotype. A common error is to test the treatment in each genotype separately and conclude that the genotypes respond differently because the effect is significant in one and not in the other. The difference between "significant" and "not significant" is not itself significant (Gelman and Stern 2006); a non-significant result may simply reflect low power. Such errors were found in about half of the neuroscience papers that compared two effects in this way (Nieuwenhuis et al. 2011). The correct analysis tests the interaction directly. Contrasts should be written down, in words and as coefficients, before the analysis.

7.4.4 Differential transcript usage

Differential expression of a gene is not the only kind of change. A gene can switch between isoforms without changing its total expression: **differential transcript usage** (or exon usage) asks whether the *proportions* of the isoforms differ between conditions. This requires transcript-level quantification (Section 7.3), whose estimates are uncertain when isoforms share most of their sequence. Quantifiers such as Salmon and kallisto estimate this uncertainty by bootstrapping or Gibbs sampling, and transcript-level estimates can be summarised to genes for gene-level tests, which also improves those tests when isoform usage changes (Soneson et al. 2016). A claim about isoforms should only be made when the reads, for example unique exons or junctions, can actually distinguish them.

7.4.5 Multiple testing

If 20,000 genes are tested at a significance level of 0.05, about 1,000 will be called significant even if nothing changes. Two kinds of error rates are controlled:

- The **family-wise error rate** (FWER) is the probability of at least one false positive. The **Bonferroni correction** controls it by testing each of m hypotheses at level α/m. Holm's step-down method (Holm 1979) is uniformly more powerful and equally valid. FWER control is appropriate when any false positive is costly, but it is very conservative for genome-scale data (Noble 2009).
- The **false discovery rate** (FDR) is the expected proportion of false positives among the genes called significant (Benjamini and Hochberg 1995). It is the natural quantity when the result is a list of candidates for follow-up. The **Benjamini–Hochberg** procedure sorts the m p-values, $p_{(1)} \le \dots \le p_{(m)}$, finds the largest k for which $p_{(k)} \le \frac{k}{m}\alpha$, and rejects the hypotheses $1, \dots, k$. This controls the FDR at level α when the tests are independent or positively correlated.

The FDR is a property of the whole list, and an *expected* value over repeated experiments, not a statement about individual genes. A list called at FDR 5% is expected to contain about 5% false discoveries, but it does not mean that each gene in it has a 5% chance of being a false positive: genes near the cutoff are much more likely to be false than those with very small p-values.

Storey's **q-value** is the smallest FDR at which a gene would be called significant (Storey and Tibshirani 2003). It also estimates the proportion of true null hypotheses, π_0, from the flat part of the p-value histogram, which gives more power when many genes change. Efron's empirical Bayes approach estimates the **local false discovery rate**, the probability that a particular gene is null given its test statistic (Efron and Tibshirani 2002). The histogram of p-values is a useful diagnostic: under the null hypothesis p-values are uniformly distributed, so a peak near zero indicates true signal, while other shapes indicate problems with the model.

Statistical significance is not biological importance. The p-value depends on the effect size, the variability and the number of replicates, so a large fold change in a noisy, lowly expressed gene may not be significant, while with enough replicates very small fold changes become significant. A **volcano plot**, the $-\log_{10}$ p-value against the log fold change, shows both at once. Results are often filtered by both FDR and fold change, but there is no universal cutoff (such as a two-fold change): what matters depends on the gene and the biological question. Filtering on the estimated fold change after testing also undermines the FDR guarantee; it is better to test directly against a minimal relevant fold change, which both limma and DESeq2 allow (McCarthy and Smyth 2009; Love et al. 2014).

Removing genes before testing can increase power, because fewer tests are corrected for. **Independent filtering** removes genes with too few reads to give a significant result, using a statistic, such as the mean count over all samples, that is independent of the test statistic when the null hypothesis is true (Bourgon et al. 2010). DESeq2 does this automatically. A filter that uses the condition labels, such as keeping only genes that look different between the groups, is not independent and invalidates the FDR.

E-values in sequence searching (Section 1.4) are a built-in multiple-testing correction: they count the expected number of chance hits in the whole database. Be aware also of the *hidden* forms of multiple testing: trying several analyses and reporting the one that "worked", testing many subgroups, or choosing thresholds after seeing the data. These practices are a major reason why many published findings fail to replicate (Ioannidis 2005).

7.4.6 Clustering and dimensionality reduction

Expression data are high-dimensional: each sample is a point in a space with one dimension per gene, and each gene a point in a space with one dimension per sample. **Unsupervised** methods find structure

in these data without using the sample labels.

Clustering groups genes or samples with similar profiles. It requires a distance measure, such as Euclidean distance or one minus the correlation, which groups genes that go up and down together regardless of their absolute level.

- **Hierarchical clustering** repeatedly joins the two closest clusters, as in UPGMA (Section 4.3), and produces a tree (dendrogram). The distance between clusters is defined by the *linkage*: single (closest members), complete (most distant members) or average. Eisen's heat maps of hierarchically clustered genes and samples became the iconic display of expression data (Eisen et al. 1998).
- ***k*-means** partitions the data into k clusters by alternating between assigning each point to the nearest cluster centre and recomputing the centres. The number of clusters must be chosen, and the result depends on the starting point.
- Networks of co-expressed genes, built by thresholding correlations, are an alternative to clusters, and modules of co-expressed genes can be related to traits (Langfelder and Horvath 2008).

Principal component analysis (PCA) finds the orthogonal directions of largest variance in the data, computed from the singular value decomposition of the centred data matrix (Lever et al. 2017). Plotting samples on the first two or three components shows the main sources of variation. If samples group by processing date rather than by condition, there is a batch effect. PCA is linear, so the first components keep the large-scale structure of the data, and the fraction of variance explained by each component tells how much of the data the plot shows.

t-SNE (Maaten and Hinton 2008) and **UMAP** (McInnes et al. 2018; Becht et al. 2018) are non-linear methods that place similar points close together in two dimensions. They preserve local neighbourhoods and reveal clusters, which has made them the standard way to visualise single-cell data (Section 7.6). However, distances between clusters, cluster sizes and densities in such plots are not meaningful, and the layout depends on parameters and random initialisation. Two clusters that lie close together, or are joined by a bridge of cells, are not necessarily related by lineage or development, and well-separated islands can be artefacts of the method. They are tools for visualisation, not for quantitative analysis; clustering and trajectory inference should be done on the underlying data, such as the PCA space, not on the two-dimensional embedding (Chari and Pachter 2023).

7.4.7 Choosing and validating a clustering

Clustering always finds clusters. Hierarchical clustering returns a tree and k-means returns k groups even for random noise. The result depends on choices made before any algorithm runs, so a clustering is a hypothesis to be checked, not a finding in itself (D'haeseleer 2005).

Distance measure. The distance decides what "similar" means. *Euclidean distance* between two gene profiles is dominated by the absolute level: two highly expressed genes are close even if one rises when the other falls. *Correlation distance*, $d = 1 - r$, compares the shape of the profiles: genes that go up and down together are close, whatever their level. With $1 - |r|$, anticorrelated genes are also grouped together, which may be wanted for a repressor and its targets but mixes opposite responses. Spearman correlation is less sensitive to outliers. The two families are related: for profiles over n samples that are centred and scaled to unit variance, the squared Euclidean distance is $2(n-1)(1-r)$.

Scaling. Clustering is done on log-transformed or variance-stabilised values, so that a few very highly expressed genes do not dominate. When genes are clustered by the shape of their response, each gene is standardised (centred and divided by its standard deviation). When samples are clustered, the genes with the largest variance dominate the distance, which is why the analysis is usually restricted to the

most variable genes; standardising every gene instead gives the noise of flat, unexpressed genes the same weight as real signal. Strong batch effects become clusters of their own.

Linkage. In hierarchical clustering, *single* linkage tends to chain points into long, straggling clusters and is sensitive to noise; *complete* linkage gives compact clusters of similar diameter; *average* linkage (UPGMA) lies between them and is common for expression data; *Ward's* method joins the pair of clusters that least increases the within-cluster variance and, like k-means, favours compact, spherical clusters. A dendrogram becomes clusters only when it is cut at some height, and the left–right order of its leaves is partly arbitrary, because each node can be flipped. Neighbouring rows in a clustered heat map are therefore not necessarily similar.

How many clusters? There is rarely one correct number. Internal criteria compare the tightness of clusters with their separation. The *elbow* in the within-cluster sum of squares as a function of k is often unclear. The *gap statistic* compares this sum with that of reference data without cluster structure (Tibshirani et al. 2001). The *silhouette* of a point is $s = (b - a)/\max(a, b)$, where a is its mean distance to the other members of its own cluster and b its mean distance to the members of the nearest other cluster; s lies between -1 and 1, and an average close to zero means that the data have little cluster structure (Rousseeuw 1987). Model-based clustering with Gaussian mixtures can choose k with a criterion such as BIC.

Validation. *Stability* is tested by resampling: the samples or genes are subsampled or bootstrapped, the data re-clustered, and pairs of items that are grouped together consistently indicate robust clusters. Consensus clustering summarises this as a matrix of co-clustering frequencies (Monti et al. 2003). Stable clusters can still be artefacts, for example of batch. *External* validation asks whether the clusters mean something biologically: whether gene clusters are enriched in functional annotation such as GO terms (Gibbons and Roth 2002) (but see the caveats in Section 7.5), or whether sample clusters agree with known labels, measured for example with the adjusted Rand index. The strongest test is that the same clusters appear in an independent data set.

Biclustering. Ordinary clustering groups genes by their behaviour across *all* samples, but many genes are co-regulated only under some conditions or in some tumours. **Biclustering** searches for submatrices: subsets of genes with a coherent pattern over subsets of samples. It was introduced to expression data by Cheng and Church in 2000, and many algorithms followed (Madeira and Oliveira 2004). Matrix factorisations, such as non-negative matrix factorisation, address the same problem by letting a gene contribute to several overlapping programs.

7.4.8 Classification with many genes and few samples

Clustering is *class discovery*; predicting a known class, such as tumour type, response to a drug or survival, is *class prediction*, a supervised learning problem (4). The early leukaemia study did both (Golub et al. 1999). Expression data are extreme for supervised learning: the number of features p (about 20,000 genes) is far larger than the number of samples n (often tens to a few hundred), the $p \gg n$ **problem**. This has several consequences:

- With more genes than samples, a linear classifier can separate *any* labelling of the training samples perfectly, including random labels. Accuracy on the training data means nothing.
- Among thousands of genes, many correlate with the outcome by chance.
- Many different gene sets predict equally well, so a signature is not unique. Most random sets of genes are significantly associated with breast cancer outcome, because so many genes correlate with proliferation (Venet et al. 2011).

Models for this situation are simple and **regularised**: nearest shrunken centroids shrinks each class mean towards the overall mean, and genes for which every class mean is shrunk all the way to it drop out

of the classifier (Tibshirani et al. 2002); penalised regression (lasso, elastic net) and linear discriminant analysis with a diagonal covariance matrix are also common. How the penalties work is described in 4.

The most common error is **selection bias**. If the genes that best separate the classes are chosen using all samples, and a classifier on those genes is then cross-validated, the test samples have already influenced the model, and the estimated error can be close to zero even for random labels (Ambroise and McLachlan 2002; Simon et al. 2003). Gene selection, and the tuning of any penalty, are part of the model and must be repeated inside each cross-validation fold (nested cross-validation). Even a correctly cross-validated estimate comes from the same cohort, laboratory and platform. A clinical signature needs validation in an independent cohort, and must improve on the established clinical predictors, such as tumour grade and size, not only on chance.

See also: further explanation in Section D.7.3; related Wikipedia articles in Section E.7.3.

Sources

Parts of this section are adapted from the Wikipedia articles Negative binomial distribution, Multiple comparisons problem, False discovery rate, Hierarchical clustering, K-means clustering, Principal component analysis and T-distributed stochastic neighbor embedding (CC BY-SA 4.0, retrieved 4 October 2026). The statistical background is adapted from Bayes' theorem, Student's t-test and p-value. Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Gene-Expression Analysis" (CC BY 4.0), used with permission.

7.5 From gene lists to biology

A differential expression analysis typically ends with a list of hundreds or thousands of genes. Interpreting such a list gene by gene is impossible, and individual genes are often not significant even when a whole pathway changes modestly. This section describes how gene lists are interpreted with prior knowledge, and how mRNA levels relate to protein levels.

7.5.1 What a change in expression means

A difference in the measured amount of a transcript does not by itself show that its transcription has changed. The amount of an RNA, M, is a balance between synthesis and degradation,

$$\frac{dM}{dt} = k_{\text{syn}} - k_{\text{deg}} M,$$

so that at steady state $M = k_{\text{syn}}/k_{\text{deg}}$. A higher level can result from faster transcription, from changes in processing or export, or from slower degradation, and an ordinary RNA-seq experiment cannot tell these apart. Methods that label newly made RNA, or that compare unspliced and spliced reads (as RNA velocity does; Section 7.6), are needed for that.

A bulk sample is also a **mixture of cell types**. The measured expression of gene g is approximately

$$A_g = \sum_c \pi_c \, a_{gc},$$

where π_c is the contribution of cell type c and a_{gc} the expression of the gene in that cell type. A_g can change because the expression within a cell type changes, or because the proportions of the cell types change, for example when a disease leads to an influx of immune cells into a tissue. These are different biological conclusions. Bulk data can be **deconvolved** into estimated proportions of cell types, using reference expression profiles of purified cells or single-cell data; CIBERSORT, for example, estimates the proportions of immune cell types from tissue expression profiles with support vector regression (Newman et al. 2015). Deconvolution depends on the reference being appropriate for the tissue. Single-cell data separate the two kinds of change directly (Section 7.6).

Finally, an expression study is usually exploratory, and its key results should be **validated** with independent samples or with a different method, such as quantitative PCR, in situ hybridisation, protein measurements or perturbation experiments. Gene set analysis summarises the same data in another way and is not an independent confirmation. Agreement between RNA and protein strengthens an association, but neither shows that the gene causes the phenotype.

7.5.2 Gene set analysis

Gene sets are groups of genes that share a function, a location or a regulatory mechanism: Gene Ontology terms (Section 2.6), pathways from KEGG or Reactome, targets of a transcription factor or microRNA, or signatures from earlier experiments. The Molecular Signatures Database (MSigDB) collects many such sets, including a small curated "hallmark" collection (Liberzon et al. 2015).

Over-representation analysis (ORA) asks whether a gene set is over-represented among the significant genes. With N genes in total, K of which belong to the set, and n significant genes, of which k are in the set, the probability of seeing k or more by chance is given by the **hypergeometric distribution** (equivalent to a one-sided Fisher's exact test):

$$P(X \geq k) = \sum_{i \geq k} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}.$$

Tools such as DAVID made this routine for many annotation types (Huang et al. 2008). ORA is simple, but it ignores the size of the changes and assumes that genes are independent. Two choices are part of the null hypothesis itself, not technical details. The first is the **background** (the universe of N genes): it should be the genes that were actually measured and could have been detected, not the whole genome. With the whole genome as background, any set of genes expressed in the tissue studied looks enriched, simply because unexpressed genes could never have been selected. The second is the **threshold** that defines the gene list: changing the FDR or fold-change cutoff changes n and k, and can change which sets are significant. Results that hold only for one background or one threshold should be treated with caution. Since thousands of gene sets are tested, multiple-testing correction is needed (Section 7.4), and because gene sets overlap heavily, the significant sets are often redundant.

Gene set enrichment analysis (GSEA) avoids the cutoff (Subramanian et al. 2005). All genes are ranked by a statistic of differential expression. For each gene set, a running-sum statistic walks down the ranked list, increasing when it meets a gene in the set and decreasing otherwise. The maximum deviation from zero, the *enrichment score*, is large when the members of the set are concentrated at the top or bottom of the list, similar to a Kolmogorov–Smirnov statistic. Significance is assessed by permuting the sample labels, which keeps the correlation between genes, or the gene labels, which is faster but ignores it. GSEA can detect pathways in which many genes change modestly, the typical situation in biology.

Methods differ in the null hypothesis they test. *Competitive* tests ask whether the genes in a set are more differentially expressed than other genes. *Self-contained* tests ask whether any of the genes in the

set are differentially expressed at all. Pathway methods that also use the topology of the network, and network-based approaches, are discussed in Chapter 11.

7.5.3 Co-expression and regulatory variation

Expression data also reveal relationships between genes. Genes with correlated expression across many conditions often share functions, which is used to predict function (Chapter 10) and to define modules (Langfelder and Horvath 2008). Combined with genotypes, expression becomes a quantitative trait: **expression quantitative trait loci** (eQTLs) are genetic variants associated with the expression of a gene. The GTEx project mapped eQTLs in about 50 tissues from nearly a thousand donors and found that almost all genes have genetic variants that affect their expression (The GTEx Consortium 2020). eQTLs are an important tool for interpreting the non-coding variants found in genome-wide association studies.

7.5.4 Protein-level expression

What matters for most cellular functions is the amount of protein, not mRNA. How well do mRNA levels predict protein levels? Early studies found correlations that were surprisingly low. A large study in mouse cells measured mRNA and protein levels and turnover for thousands of genes, and concluded that protein levels are determined mainly by translation rates (Schwanhäusser et al. 2011). A later reanalysis argued that much of the apparent discrepancy is due to measurement noise, and that across genes in steady state, mRNA levels explain most of the variation in protein levels, while during dynamic changes, and for comparisons of the same gene between conditions, post-transcriptional regulation is more important (Liu et al. 2016).

The **Human Protein Atlas** (HPA), a Swedish project started in 2003, combines RNA-seq of human tissues and cell lines with antibody-based imaging of the corresponding proteins, and classifies each gene by where it is expressed, from ubiquitous *housekeeping* genes to genes enriched in a single tissue (Uhlén et al. 2015). Mass-spectrometry and antibody-based measurement of proteins are described in Chapter 8, and the HPA in Section 2.8.

See also: further explanation in Section D.7.4; related Wikipedia articles in Section E.7.4.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Gene set enrichment analysis, Hypergeometric distribution, Expression quantitative trait loci and Human Protein Atlas (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Gene-Expression Analysis" (CC BY 4.0), used with permission.

7.6 Single-cell genomics

Bulk RNA-seq measures the average expression of millions of cells. A tissue is a mixture of cell types, and even cells of the same type differ, so the average can hide the most interesting biology. **Single-cell RNA-seq** (scRNA-seq) measures the transcriptome of each cell separately. It has become one of the fastest-developing areas of genomics, with its own experimental methods, statistical models and analysis workflows (Luecken and Theis 2019; Heumos et al. 2023).

7.6.1 Technologies

The first single-cell transcriptome was published in 2009 (Tang et al. 2009). The number of cells that can be profiled in one experiment has since grown exponentially, from a handful to millions (Svensson et al. 2018):

- **Plate-based methods**, such as Smart-seq2 (Picelli et al. 2013), sort single cells into wells and produce full-length cDNA. They give high sensitivity and full transcript coverage for hundreds to thousands of cells.
- **Droplet-based methods**, such as Drop-seq (Macosko et al. 2015), inDrop (Klein et al. 2015) and the commercial 10x Genomics Chromium platform (Zheng et al. 2017), encapsulate single cells in nanolitre droplets together with beads carrying barcoded primers. Every transcript from the same cell gets the same **cell barcode**, so tens of thousands of cells can be sequenced together and separated computationally.
- **Combinatorial indexing** methods label cells in several rounds of split-pool barcoding without physically isolating them, and reach millions of cells.

Most high-throughput methods sequence only the 3' or 5' end of each transcript, and tag each original molecule with a **unique molecular identifier** (UMI), a random sequence added before amplification (Islam et al. 2013). Reads with the same UMI come from the same molecule, so counting UMIs instead of reads removes amplification bias.

7.6.2 Properties of the data

A typical cell contains a few hundred thousand mRNA molecules, of which only a few thousand to tens of thousands are captured. The data are therefore a very **sparse** matrix of counts, with most entries zero. Many zeros were first thought to be technical *dropouts* requiring special zero-inflated models, but for UMI data the number of zeros is largely what is expected from sampling a small number of molecules at the given sequencing depth, and negative binomial or Poisson models fit well (Svensson 2020). Whether extra zeros need to be modelled therefore depends on the protocol and the model; it may matter more for read counts from full-length, non-UMI protocols. Many observed zeros also reflect real biological differences between cell types. Other technical problems are specific to single cells: empty droplets, *doublets* (two cells with the same barcode), ambient RNA from lysed cells, and dying cells with high fractions of mitochondrial reads.

7.6.3 The analysis workflow

The standard workflow, implemented in the R package **Seurat** (Hao et al. 2021) and the Python package **Scanpy** (Wolf et al. 2018), consists of (Luecken and Theis 2019; Heumos et al. 2023):

1. **Quality control**: removing barcodes with too few genes, too many counts (possible doublets) or a high mitochondrial fraction.
2. **Normalisation** for the number of molecules captured per cell, for example by scaling and log-transforming, by pooling cells to estimate size factors (Lun et al. 2016), or with regularised negative binomial regression (Hafemeister and Satija 2019).
3. **Feature selection**: keeping the most variable genes.
4. **Dimensionality reduction** with PCA (Section 7.4).
5. Building a ***k*-nearest-neighbour graph** of cells in PCA space and **clustering** it with community-detection algorithms such as Louvain or Leiden (Traag et al. 2019).
6. **Visualisation** with UMAP or t-SNE.

7. Finding **marker genes** for each cluster and **annotating** the clusters as cell types, by comparison with known markers or by transferring labels from annotated reference atlases.

A cluster is a computational result, not automatically a cell type. The number of clusters depends on the resolution parameter, on the choice of genes and on batch correction; a continuous process such as differentiation can be cut into several arbitrary clusters; and residual batch effects, doublets or stressed cells can form clusters of their own. Annotation is therefore an inference that should be supported by several known markers, by biological context and, where possible, by independent validation, for example with imaging or protein markers.

Further analyses depend on the question:

- **Data integration.** Samples processed separately differ by batch effects, which must be removed before cells from different samples, individuals or technologies can be compared, without removing real biological differences. Harmony iteratively corrects the PCA embedding (Korsunsky et al. 2019). scVI uses a variational autoencoder, a deep generative model of the counts with batch as a covariate (Lopez et al. 2018). Integration is a trade-off between mixing the batches and keeping the biological variation, and a large benchmark found that the balance depends on the method and the preprocessing, in particular when biology and batch overlap (Luecken et al. 2021). If a disease state is present in only one batch, forcing the batches to mix can remove the disease signal. An integrated embedding is useful for clustering, visualisation and transfer of labels, but tests between conditions should go back to the uncorrected counts and a design that respects donors (Figure 7.3).
- **Trajectory inference.** Cells in development or differentiation form continua rather than discrete clusters. *Pseudotime* methods order cells along a path through expression space (Trapnell et al. 2014). **RNA velocity** uses the ratio of unspliced to spliced reads to estimate whether each gene is being switched on or off, and so predicts the future state of each cell (La Manno et al. 2018).
- **Differential expression** between conditions should be tested between *samples*, not between individual cells. Cells from the same individual share its genotype, environment and sample processing, so they are not independent biological replicates; the experimental unit is the individual (Section 7.3). Treating thousands of cells as replicates is pseudoreplication and gives very many false positives (Squair et al. 2021; Zimmerman et al. 2021). In **pseudobulk** analysis, the counts of all cells of one cell type are summed for each sample, and the resulting sample-level counts are analysed with bulk methods such as edgeR, DESeq2 or limma. Mixed models with the individual as a random effect are an alternative; a comparison of the two strategies found that pseudobulk methods performed well (Crowell et al. 2020). More cells improve the description of each individual, but only more individuals support conclusions about a population.
- **Differential abundance** asks a different question: whether the *proportions* of cell types or states differ between conditions, for example an expansion of a T-cell population in disease, rather than whether genes change within a cell type. Proportions are compositional, since an increase in one cell type must lower the share of the others, and they must also be compared between samples, not between cells. Methods test either the cluster proportions or local neighbourhoods in the kNN graph, which avoids depending on a fixed clustering (Dann et al. 2021).

7.6.4 Multimodal single-cell data

RNA is only one layer of a cell's state. **Multimodal** (*multiome*) methods measure more than one layer in the same cell, using the shared cell barcode to link them. CITE-seq adds oligonucleotide-tagged antibodies, so that surface proteins are counted together with the transcriptome (Stoeckius et al. 2017). Joint RNA and ATAC methods, such as the 10x Multiome kit, measure gene expression and chromatin accessibility (Section 6.9) in the same nucleus, which links candidate regulatory regions to the genes whose expression they correlate with. Because each modality has its own noise and sparsity, they

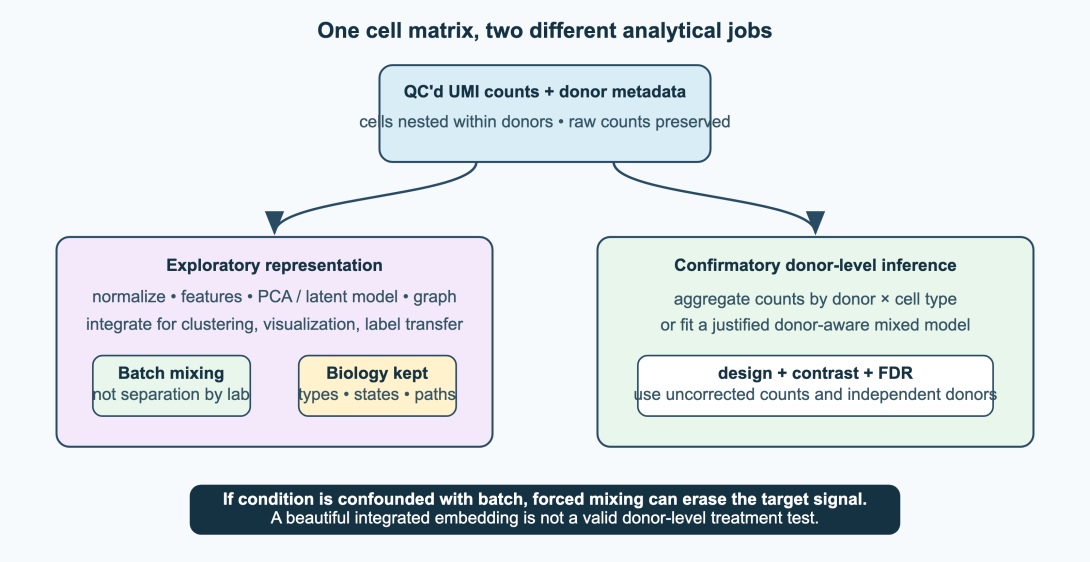


Figure 7.3: One single-cell count matrix supports two different analyses. Exploratory analyses of an integrated representation are used for clustering, visualisation and label transfer; tests between conditions aggregate the uncorrected counts by donor and cell type, or use mixed models with the donor as a random effect. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

are first analysed separately and then combined, for example with Seurat's *weighted nearest neighbour* method, which learns for each cell how much to trust each modality (Hao et al. 2021). When modalities are measured in different cells, they can still be integrated by mapping both into a common space, but the link between them is then inferred rather than observed.

7.6.5 Spatial transcriptomics

Dissociating a tissue into single cells loses the information about where each cell was. **Spatial transcriptomics** measures expression while keeping its position, and was named Method of the Year 2020 by *Nature Methods* (Marx 2021). There are two families of methods:

- **Sequencing-based methods** capture RNA from a tissue section on an array of spatially barcoded spots. The method was developed in Stockholm by Patrik Ståhl, Joakim Lundeberg and colleagues (Ståhl et al. 2016) and commercialised as 10x Visium. Early spots were 100 µm across and contained several cells. Later methods, such as Slide-seq, reach near-cellular resolution (Rodriques et al. 2019).
- **Imaging-based methods**, such as MERFISH, detect individual RNA molecules in place with multiple rounds of fluorescent hybridisation, and measure hundreds to thousands of chosen genes at subcellular resolution (Chen et al. 2015).

Analysis tasks include identifying spatially variable genes and tissue domains, *deconvolution* of spots into mixtures of cell types using single-cell references, and the study of cell–cell communication between neighbouring cells. Spatial data bring their own statistical problems. Neighbouring spots are correlated (*spatial autocorrelation*), so methods that treat locations as independent overstate the evidence; tissue edges, histology and local cell density can mimic spatial patterns; and a single section is one sample, so conclusions about a tissue need replicate sections, ideally from several individuals. Platforms also differ in sensitivity as well as resolution, and the nominal spot size alone is not a sufficient basis for choosing between them (You et al. 2024). Spatial and multimodal methods are converging: spatial measurements of proteins and chromatin accessibility are being added to spatial RNA, and dissociated

single-cell multiome data are mapped onto tissue sections, so that each cell's type, regulatory state and position can be studied together.

7.6.6 Foundation models for single-cell data

Following the success of language models for proteins (Section 1.10), **foundation models** have been trained on tens of millions of single-cell transcriptomes. Geneformer represents each cell as a list of its genes ranked by expression, and learns with masked-gene prediction (Theodoris et al. 2023). scGPT uses a generative transformer (Cui et al. 2024). The models produce embeddings of genes and cells that can be fine-tuned for cell-type annotation, batch integration or the prediction of the effects of perturbations. Their value is still debated: independent evaluations found that, without fine-tuning, their embeddings often do not outperform simple baselines such as highly variable genes and PCA (Kedzierska et al. 2025), and that for predicting the effects of genetic perturbations they do not yet beat simple linear models (Ahlmann-Eltze et al. 2025). This is an area where students should expect rapid change, and where critical evaluation against simple baselines is essential (Chapter 12).

7.6.7 Cell types and cell atlases

Single-cell data have revived an old question: **what is a cell type?** Classically, cell types were defined by morphology, location, function or a few marker proteins. Single-cell transcriptomics proposes them as clusters in expression space, which raises new questions: at which resolution should clusters be cut, how do discrete types relate to continuous states, and how should types defined by expression be matched to those defined by other properties (Zeng 2022)? The **Human Cell Atlas** aims to map all cell types in the human body (Regev et al. 2017), and atlases of whole organisms, such as Tabula Muris (The Tabula Muris Consortium 2018) and Tabula Sapiens (The Tabula Sapiens Consortium 2022), provide references across organs. Such references, together with ontologies of cell types, make it possible to annotate new data sets automatically. The atlases, and the efforts to build models of cells from them, are discussed further in Section 13.7.

See also: further explanation in Section D.7.5; related Wikipedia articles in Section E.7.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Single-cell transcriptomics, Single-cell sequencing, Unique molecular identifier, Spatial transcriptomics, RNA velocity, Foundation model and Human Cell Atlas (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Gene-Expression Analysis" (CC BY 4.0), used with permission.

7.7 Single-cell sequencing and cancer evolution

Cancer is an evolutionary process within the body. Single-cell methods have made it possible to follow this process cell by cell, and to apply to tumours the phylogenetic methods developed for species (Chapter 4).

7.7.1 Cancer as clonal evolution

In 1976 Peter Nowell proposed that tumours evolve like populations of organisms: they start from a single cell, and successive mutations followed by selection of the fittest subclones lead to increasingly malignant cell populations (Nowell 1976). The concepts of Chapter 3 apply directly: mutation, selection, drift and population size all shape a tumour, and the result is **intratumour heterogeneity**, genetically different subclones within the same tumour. This heterogeneity matters clinically, because a treatment that kills the dominant clone selects for resistant subclones, which then cause relapse.

Bulk sequencing of tumours measures the frequency of each mutation in a mixture of normal and tumour cells. *Clonal* mutations, present in all tumour cells, occurred early, while *subclonal* mutations, present in only some, occurred later. Sequencing several regions of the same tumour showed that a biopsy from one region misses many of the mutations, and that tumours evolve in a **branched** rather than a linear way (Gerlinger et al. 2012; McGranahan and Swanton 2017). Reconstructing the subclones and their phylogeny from bulk data is a deconvolution problem with many possible solutions (Schwartz and Schäffer 2017).

7.7.2 Single-cell DNA sequencing

Single-cell sequencing removes the need for deconvolution. A cell contains only two copies of its genome, about 6 pg of DNA, which must be amplified before sequencing, and whole-genome amplification introduces errors, uneven coverage and the loss of one allele (*allelic dropout*) (Gawad et al. 2016). Copy-number changes, which affect large regions, can nevertheless be measured robustly even with very shallow sequencing of each cell. Navin and colleagues used single-nucleus sequencing of breast tumours to show that copy-number profiles grouped cells into a few major clones, and that a tumour can arise through a short burst of large genomic changes followed by stable clonal expansion (*punctuated* evolution) (Navin et al. 2011). High-throughput single-cell DNA sequencing of thousands of cells later showed that breast tumours maintain a reservoir of subclonal diversity during their growth (Minussi et al. 2021).

Trees of single cells are built with methods adapted from phylogenetics (Section 4.3), with models that account for the specific errors of single-cell data, such as allelic dropout and false-positive mutations, and for the fact that copy-number changes are not independent events at single sites (Schwartz and Schäffer 2017). Many mutation-tree methods rest on the **infinite-sites assumption**: each site mutates at most once in the history of the tumour, and a mutation, once gained, is never lost. Cancer genomes violate this. A deletion or a loss of heterozygosity can remove a mutation from a subclone, and copy-number changes alter the number of copies of a mutated allele, which also distorts the variant allele frequencies used in bulk deconvolution. In addition, a single biopsy or a small sample of cells may miss subclones altogether. Trees inferred under the infinite-sites model should therefore be read as hypotheses, and newer methods relax the assumption, for example by allowing mutation losses or by modelling single-nucleotide and copy-number changes jointly.

7.7.3 Single-cell transcriptomics of tumours

Single-cell RNA-seq of tumours measures both the cancer cells and the **tumour microenvironment**: immune cells, fibroblasts and blood vessels (Lawson et al. 2018). In glioblastoma, cells from the same tumour expressed programmes characteristic of different molecular subtypes, which had been defined from bulk data as properties of whole tumours (Patel et al. 2014). In metastatic melanoma, scRNA-seq separated malignant from non-malignant cells, identified cancer cell states associated with drug resistance and characterised the state of the infiltrating T cells (Tirosh et al. 2016).

Cancer cells are usually identified in scRNA-seq data by inferring large **copy-number changes** from the expression of neighbouring genes along each chromosome, since an amplified region raises the average expression of all its genes (Patel et al. 2014; Gao et al. 2021). This also makes it possible to group cancer cells into genetic subclones and link genotype to transcriptional state.

Single-cell studies have shown that not all heterogeneity is genetic. Cancer cells with the same genome can occupy different transcriptional or epigenetic states, and switch between them. Some of these states, such as slowly dividing *persister* cells, survive treatment without any resistance mutation. Integrating genetic and non-genetic heterogeneity, using multi-omic methods that measure DNA, RNA and chromatin in the same cell, and lineage tracing, is a major current goal (Nam et al. 2020).

See also: further explanation in Section D.7.6; related Wikipedia articles in Section E.7.6.

Sources

Parts of this section are adapted from the Wikipedia articles Somatic evolution in cancer, Tumour heterogeneity and Single-cell sequencing (CC BY-SA 4.0, retrieved 4 October 2026).

7.8 Key concepts

A checklist for revision.

- **Before microarrays:** Northern blots, ESTs and SAGE
- **Microarrays:** spotted cDNA (two-colour) vs. Affymetrix oligonucleotide (one-colour) arrays; Illumina BeadArrays and SNP genotyping arrays; background correction, log transformation, MA plots and loess, quantile normalisation, RMA; MIAME and reproducibility; limitations (known sequences, cross-hybridisation, non-linear intensity response and dynamic range)
- **Landmark microarray studies:** co-expression clusters; cancer classification and breast cancer subtypes; expression signatures in the clinic; tissue atlases
- **RNA-seq experiments:** poly(A) selection vs. rRNA depletion; strand-specific libraries; sequencing depth; the experimental unit; biological vs. technical replicates, pseudoreplication and how many replicates are needed; batch effects and confounding; randomisation, blocking and paired designs
- **RNA-seq quality control:** RNA-specific statistics (unique mapping, rRNA and mitochondrial fractions, strandedness, gene-body coverage and 3′ bias); sample-level PCA and identity checks; outliers as a reason to investigate; the SEQC/MAQC-III reproducibility study
- **RNA-seq analysis:** spliced alignment (TopHat, STAR, HISAT2); reference-guided and de novo transcript assembly (Cufflinks, StringTie, Trinity); quantification with EM (RSEM); pseudo-alignment and equivalence classes (kallisto, Salmon); long-read RNA-seq for full-length isoforms and transcript discovery
- **Normalisation:** CPM, RPKM/FPKM, TPM and what TPM does not correct; compositional data and composition bias; TMM and median-of-ratios size factors; global shifts and spike-ins; raw counts as input to DE tools
- **Differential expression:** why a plain t-test fails; empirical Bayes and moderated statistics (limma, voom); Poisson vs. negative binomial models; dispersion and its shrinkage; GLMs and design matrices; fold-change shrinkage (DESeq2, edgeR); Wald, likelihood-ratio and quasi-likelihood F-tests
- **Statistical background:** Bayes' theorem and base rates; binomial, Poisson, normal, negative binomial and extreme value distributions; likelihood; p-values and what they are not; effect sizes and confidence intervals; t-tests, non-parametric, χ^2/Fisher and permutation tests; correlation; type I/II errors and power

- **Design and contrasts:** design matrices and full rank; paired designs (`~ donor + treatment`); contrasts; interactions as a difference of differences; why "significant in one group but not the other" is not a test of a difference
- **Differential transcript usage:** isoform proportions vs. total gene expression; uncertainty of transcript estimates; summarising transcripts to genes
- **Multiple testing:** FWER and Bonferroni; FDR and the Benjamini–Hochberg procedure; q-values and π_0; local FDR; why FDR is not a per-gene probability; p-value histograms; significance vs. effect size; volcano plots and fold-change thresholds; independent filtering
- **Clustering and dimensionality reduction:** distance measures; hierarchical clustering and linkage; k-means; co-expression networks; PCA; t-SNE and UMAP and their pitfalls (distances and proximity are not lineage)
- **Choosing and validating clusters:** clustering always finds clusters; Euclidean vs. correlation distance and what each groups together; log transformation, standardising genes and selecting variable genes; single, complete, average and Ward linkage; cutting a dendrogram; choosing k (elbow, gap statistic, silhouette, BIC); stability by resampling and consensus clustering; external validation with annotation and independent data; biclustering
- **Classification from expression data:** class discovery vs. class prediction; the $p \gg n$ problem and perfect separation of random labels; unstable signatures and random signatures; nearest shrunken centroids and penalised regression; selection bias when genes are chosen outside cross-validation; independent validation cohorts and clinical baselines
- **What an expression change means:** synthesis and degradation at steady state; bulk samples as mixtures of cell types; changes in composition vs. changes within cells; deconvolution (CIBERSORT); validation with independent samples and orthogonal methods
- **Interpretation:** gene sets and MSigDB; over-representation analysis and the hypergeometric test; choice of background and threshold; GSEA; competitive vs. self-contained tests; eQTLs and GTEx
- **mRNA and protein:** how well mRNA predicts protein levels; the Human Protein Atlas
- **Single-cell RNA-seq:** plate-based vs. droplet methods; cell barcodes and UMIs; sparsity and whether zeros are inflated; doublets, ambient RNA and QC; normalisation, HVGs, PCA, kNN graphs, Leiden clustering, UMAP; marker genes and annotation, and why clusters are not automatically cell types; integration (Harmony, scVI) and the trade-off between batch mixing and conserved biology; pseudotime and RNA velocity; pseudobulk differential expression and why cells are not replicates; differential abundance of cell types; multimodal data (CITE-seq, joint RNA+ATAC, weighted nearest neighbours)
- **Spatial transcriptomics:** sequencing-based (spatial barcodes, Visium, Slide-seq) vs. imaging-based (MERFISH); deconvolution; spatial autocorrelation and replicate sections
- **Foundation models:** Geneformer and scGPT; evaluations against simple baselines
- **Cell types:** what is a cell type; the Human Cell Atlas, Tabula Muris and Tabula Sapiens
- **Cancer evolution:** Nowell's clonal evolution; intratumour heterogeneity; clonal vs. subclonal mutations; branched and punctuated evolution; single-cell DNA sequencing and its errors; the infinite-sites assumption and how copy-number changes and loss of heterozygosity violate it; inferring copy number from scRNA-seq; tumour microenvironment; non-genetic heterogeneity and persister cells

7.9 Recommended reading

These papers are recommended for reading in full:

- Wang et al. (2009) — RNA-seq, a revolutionary tool for transcriptomics; why sequencing replaced arrays.

- Conesa et al. (2016) — a survey of best practices for RNA-seq data analysis, from design to interpretation.
- Love et al. (2014) — DESeq2; negative binomial models, dispersion and fold-change shrinkage.
- Benjamini and Hochberg (1995) — the false discovery rate; the most important idea in genome-scale statistics.
- Luecken and Theis (2019) — best practices in single-cell RNA-seq analysis, step by step.
- Nam et al. (2020) — cancer evolution studied with single-cell multi-omics.

Titles, abstracts, links and the full texts are in Section A.7; a longer list of further papers is in Section B.7.

7.10 Background reading (open access)

- **DNA microarray (Wikipedia)** — A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles of a specific DNA sequence, known as probes.
  https://en.wikipedia.org/wiki/DNA_microarray
- **RNA-Seq (Wikipedia)** — RNA-Seq is a next-generation sequencing (NGS) technique used to quantify and identify RNA molecules in a biological sample, providing a snapshot of the transcriptome at a specific time. It enables transcriptome-wide analysis by sequencing cDNA derived from RNA.
  https://en.wikipedia.org/wiki/RNA-Seq
- **Transcriptomics technologies (Wikipedia)** — Transcriptomics technologies are the techniques used to study an organism's transcriptome, the sum of all of its RNA transcripts. The information content of an organism is recorded in the DNA of its genome and expressed through transcription.
  https://en.wikipedia.org/wiki/Transcriptomics_technologies
- **Multiple comparisons problem (Wikipedia)** — Multiple comparisons, multiplicity or multiple testing problem occurs when many statistical tests are performed on the same dataset. Each test has its own chance of a Type I error, so the overall probability of making at least one false positive increases as the number of tests grows.
  https://en.wikipedia.org/wiki/Multiple_comparisons_problem
- **Bonferroni correction (Wikipedia)** — Bonferroni correction is a method to counteract the multiple comparisons problem in statistics. It is named after the mathematician Carlo Emilio Bonferroni.
  https://en.wikipedia.org/wiki/Bonferroni_correction
- **False discovery rate (Wikipedia)** — In statistics, the false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the FDR, which is the expected proportion of "discoveries" that are false.
  https://en.wikipedia.org/wiki/False_discovery_rate
- **Bayes' theorem (Wikipedia)** — In probability theory, Bayes' theorem, named after Thomas Bayes, gives a mathematical rule for inverting conditional probabilities, allowing the probability of a cause to be found given its effect.
  https://en.wikipedia.org/wiki/Bayes%27_theorem
- **Student's t-test (Wikipedia)** — Student's t-test is a statistical test used to test whether the difference between the response of two groups is statistically significant or not. It is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.
  https://en.wikipedia.org/wiki/Student%27s_t-test

- **p-value (Wikipedia)** — In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.
  https://en.wikipedia.org/wiki/P-value
- **Statistics for Biologists** — Nature's collection of short articles for biologists, including the *Points of Significance* column from Nature Methods, explaining p-values, power, multiple testing, regression, PCA and other statistical concepts.
  https://www.nature.com/collections/qghhqm
- **Think Stats** — Allen Downey's free book introducing probability and statistics through Python programming and exploratory data analysis.
  https://greenteapress.com/wp/think-stats-3e/
- **Principal component analysis (Wikipedia)** — Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.
  https://en.wikipedia.org/wiki/Principal_component_analysis
- **Gene set enrichment analysis (Wikipedia)** — Gene set enrichment analysis (GSEA) (also called functional enrichment analysis or pathway enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with different phenotypes (e.g. different organism growth patterns or diseases).
  https://en.wikipedia.org/wiki/Gene_set_enrichment_analysis
- **Single-cell transcriptomics (Wikipedia)** — Single-cell transcriptomics refers to the quantification and analysis of the transcriptomes of individual cells. Single-cell transcriptomics makes it possible to unravel heterogeneous cell populations, reconstruct cellular developmental pathways, and model transcriptional dynamics.
  https://en.wikipedia.org/wiki/Single-cell_transcriptomics
- **Spatial transcriptomics (Wikipedia)** — Spatial transcriptomics, or spatially resolved transcriptomics, is a method that captures positional context of transcriptional activity within intact tissue.
  https://en.wikipedia.org/wiki/Spatial_transcriptomics
- **Somatic evolution in cancer (Wikipedia)** — Somatic evolution is the accumulation of mutations and epimutations in somatic cells during a lifetime, and the effects of those mutations and epimutations on the fitness of those cells. This evolutionary process has first been shown by the studies of Bert Vogelstein in colon cancer.
  https://en.wikipedia.org/wiki/Somatic_evolution_in_cancer

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

7.11 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

7.12 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Gene-Expression Analysis" and "Statistics and Machine Learning for Bioinformatics" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

7.12.1 Designing a bulk RNA-seq experiment

A study aims to identify genes affected by a drug in cultures derived from patients. Formulate an experimental and statistical design that supports a conclusion about the patient population.

Extensions, solution outline and common errors

Extensions:

- What is a biological replicate, and is a technical replicate useful?
- Would you prefer more samples or more sequencing depth?
- How do you handle paired control and treated cultures from each patient?
- What if all controls were prepared in one batch and all treated samples in another?

Solution outline:

1. Define what is to be estimated: the average change associated with the drug within patients, in the target population.
2. Treat patients, not wells or reads, as the independent experimental units; base the sample size on the expected effect, the variability and the number of tests.
3. Pair treatment within patients, randomise and balance the processing, and decide the exclusion rules and the primary contrast in advance.
4. Use a design such as `~ patient + treatment`; add nuisance covariates that can be estimated, and an interaction only if the question requires it.
5. Once the coverage is adequate, prefer more biological replicates to deeper sequencing; technical replicates describe the measurement, not the population.
6. State that perfect confounding of treatment and batch cannot be repaired computationally.

Common errors:

- Calling several wells from one culture independent biological replicates.
- Choosing the number of samples from "typical RNA-seq practice" without considering the effect size or the dispersion.
- Promising that batch correction will recover a perfectly confounded treatment effect.
- Letting a PCA plot decide the design after the results are visible.

Connections: experimental design and replication (Section 7.3); power and statistical background (Section 7.4.1); sequencing technologies (Section 5.2).

7.12.2 Alignment versus pseudo-alignment

Decide whether RNA-seq reads should be aligned to a genome or assigned to a transcriptome by pseudo-alignment or lightweight mapping, for a specified biological analysis.

Extensions, solution outline and common errors

Extensions:

- What does pseudo-alignment keep, and what does it discard?
- How are reads compatible with several isoforms assigned?
- When is genome alignment necessary?
- What happens if the annotation is incomplete?

Solution outline:

1. Start from the feature of interest: gene abundance, transcript abundance, new splice junctions, fusion genes, allele-specific expression or evidence to inspect in a genome browser.
2. Explain spliced genome alignment as placing reads on genomic coordinates across introns, followed by counting or transcript assembly.
3. Explain equivalence classes of compatible transcripts and the probabilistic assignment of reads with a mixture likelihood and the EM algorithm.
4. Prefer fast transcriptome quantification for the abundance of annotated transcripts, and genome alignment for new features or questions that depend on coordinates.
5. Record the versions of genome, transcriptome and annotation, the strandedness and the treatment of multi-mapping reads.
6. Validate isoforms that are hard to distinguish with junction-specific or targeted evidence, and take the uncertainty of the estimates into account.

Common errors:

- Saying that pseudo-alignment means "no mapping".
- Assuming that paired-end reads uniquely identify every isoform.
- Ignoring mismatches between the reference genome and the annotation.
- Treating transcript estimates from shared reads as direct observations.

Connections: spliced alignment and quantification (Section 7.3); differential transcript usage (Section 7.4); exact matching and indexes (Section 1.4); read mapping (Section 5.4); gene annotation (Section 5.6).

7.12.3 Normalisation and composition effects

Explain why raw counts cannot be compared directly between libraries, and why testing TPM values does not in general give valid differential expression results that take replicates into account.

Extensions, solution outline and common errors

Extensions:

- What do TMM and the median-of-ratios normalisation of DESeq2 estimate?
- What is the difference between correcting for library depth and for transcript length?
- What if nearly all genes truly increase in every cell?
- Are spike-ins a complete solution?

Solution outline:

1. Counts depend on sequencing depth, library composition, feature length and protocol bias, as well as on abundance.
2. Describe TMM and median-of-ratios scaling as robust estimates of the relative depth, under the assumption that most genes do not change.
3. Include the scaling factors in the count model as offsets, and use the raw counts for inference.
4. Describe TPM as a length-adjusted proportion within a sample, useful for describing abundance but not a model of the variance between replicates.
5. Explain that a global change per cell cannot be detected from standard relative libraries without an external reference.

6. Discuss spike-ins and cell counts as additional assumptions rather than complete solutions.

Common errors:

- Equating normalised counts with absolute numbers of molecules per cell.
- Saying that normalisation makes all samples biologically identical.
- Applying batch correction or quantile normalisation before asking what signal it removes.
- Comparing RPKM or TPM values with a t-test, ignoring the dependence of the variance on the mean.

Connections: normalisation (Section 7.3); microarray normalisation (Section 7.2); relative and absolute quantification in proteomics (Section 8.4).

7.12.4 Negative binomial differential expression

Derive and interpret the negative binomial generalised linear model used by DESeq2 or edgeR for differential expression with replicates.

Extensions, solution outline and common errors

Extensions:

- Why not use a Poisson model?
- What is dispersion, and why is it shrunk?
- How do Wald, likelihood-ratio and quasi-likelihood tests differ?
- How should fold changes and the FDR be reported?

Solution outline:

1. Write $Y_{gi} \sim \mathrm{NB}(\mu_{gi}, \alpha_g)$ and $\mathrm{Var}(Y) = \mu + \alpha\mu^2$.
2. Write $\log \mu_{gi} = \log s_i + x_i^T \beta_g$, and define the offset, the row of the design matrix, the coefficients and the contrast.
3. Explain that the Poisson variance equals the mean and misses the overdispersion between biological replicates.
4. Explain how noisy dispersion estimates of individual genes are shrunk towards a trend fitted across all genes (empirical Bayes).
5. Use a Wald test for a single coefficient or contrast, a likelihood-ratio test for a question involving several coefficients, or a quasi-likelihood test to account for uncertainty in the dispersion.
6. Report the effect size, its uncertainty, the adjusted p-value and the abundance; shrinkage stabilises noisy estimates but does not create evidence.

Common errors:

- Calling dispersion only technical noise.
- Interpreting a shrunken log fold change as the observed change.
- Defining the FDR as the probability that a particular gene is a false positive.
- Choosing the test that gives the longest gene list.

Connections: differential expression and multiple testing (Section 7.4); linear models (4); peptide- and protein-level FDR (Section 8.4).

7.12.5 Contrasts, interactions and batch

A drug effect is statistically significant in wild type but not in a mutant. Does this show that the effect of the drug depends on the genotype? Formulate the appropriate test.

Extensions, solution outline and common errors

Extensions:

- Which contrast tests whether the effect differs?
- What if genotype and batch are correlated?
- What makes a design matrix rank-deficient?
- How would you analyse an experiment with several time points?

Solution outline:

1. Reject the comparison of two significance labels; test a difference of differences.
2. Specify $(\text{mut}_{\text{drug}} - \text{mut}_{\text{control}}) - (\text{WT}_{\text{drug}} - \text{WT}_{\text{control}})$ in a model with main effects and an interaction.
3. Include batch when it can be estimated; inspect the design matrix, and explain that perfect confounding makes separate coefficients impossible.
4. For time courses, test the condition-by-time terms jointly, for example with a likelihood-ratio test.
5. Keep the independent experimental unit and any repeated measurements explicit.
6. Interpret an interaction as a context-dependent association, not automatically as a mechanism.

Common errors:

- "Significant here, not significant there" as a test of interaction.
- Removing a batch effect until the groups look mixed in a PCA plot.
- Including more coefficients than the design supports.
- Treating repeated time points from the same individual as independent.

Connections: design matrices, contrasts and interactions (Section 7.4); batch effects and confounding (Section 7.3); linear models (4).

7.12.6 Single-cell differential expression

A data set contains 10,000 cells from each of three treated and three control patients. What is the sample size for a conclusion about the effect of treatment in the patient population? Formulate a valid analysis of differential expression within a cell type.

Extensions, solution outline and common errors

Extensions:

- Why are cells not independent replicates?
- What is pseudobulk analysis?
- Are zeros "dropouts" that require a zero-inflated model?
- How do you separate differential abundance from differential expression within a cell type?

Solution outline:

1. The sample size is six patients; the cells are nested measurements within patients.
2. Sum the UMI counts per patient and cell type (pseudobulk) and use a bulk count model, or use a mixed model with the patient as a random effect.
3. Treat UMI zeros as non-detections; do not assume extra zero inflation without evidence.
4. Distinguish a change of expression within a cell population from a change in the frequency of that population.
5. Check that each patient contributes enough cells, and show the results per patient.
6. Distinguish exploratory marker genes from tests between conditions.

Common errors:

- Reporting a tiny p-value driven by thousands of correlated cells.
- Downsampling cells and claiming that this creates biological independence.
- Confusing more immune cells with activation within immune cells.

Connections: single-cell analysis, pseudobulk and differential abundance (Section 7.6); experimental units (Section 7.3); pitfalls in evaluation (4).

7.12.7 Clustering, annotation and integration

After integration of several single-cell RNA-seq batches, the cells appear well mixed in a low-dimensional embedding. Is this enough to show that the integration was successful?

Extensions, solution outline and common errors

Extensions:

- What should be preserved, as well as mixed?
- Is a cluster a cell type?
- Can the corrected matrix be used for differential expression?
- What if a disease state occurs in only one batch?

Solution outline:

1. Define the purpose of the integration: visualisation, clustering, mapping to a reference or label transfer.
2. Assess batch mixing and the conservation of known cell types, continuous states, rare populations and marker genes separately.
3. Explain that biology confounded with batch cannot be recovered; forced mixing may erase a disease signal.
4. Annotate clusters with several positive and negative markers, reference data and consistency across samples.
5. Keep the uncorrected counts and use models that take patients into account for tests between conditions.
6. Compare several methods and parameters; an attractive UMAP is not validation.

Common errors:

- Treating maximal batch mixing as the only goal.
- Naming a cluster after one marker gene.
- Reading distances between clusters or cluster sizes literally from a UMAP.

Connections: single-cell analysis and integration (Section 7.6); clustering and dimensionality reduction (Section 7.4); unsupervised learning (4).

7.12.8 Spatial transcriptomics and validation

Design an analysis to determine whether an immune-response programme is localised at the boundary of a tumour, and distinguish spatial association from evidence of signalling between cells.

Extensions, solution outline and common errors

Extensions:

- Would you choose sequencing-based spatial transcriptomics or an imaging-based method?
- What does deconvolution of spots infer?
- Why is ordinary differential expression between spots risky?
- How would you validate a proposed ligand–receptor interaction?

Solution outline:

1. Define the boundary, the target genes, the spatial scale, and whether the aim is transcriptome-wide discovery or targeted measurement.
2. Choose sequencing-based capture for broad discovery, or imaging for targeted measurement of single molecules, balancing field of view, sensitivity and resolution.
3. Align expression to histology, account for spots that contain several cells, segmentation errors and spatial autocorrelation.
4. Use replicate sections from several patients or animals; neighbouring spots are not independent individuals.
5. Describe deconvolution as inference of mixtures that depends on a reference, not as direct observation of single cells.
6. Validate expression with an independent imaging method, and signalling with protein localisation, perturbation or a functional response; co-expression alone is a hypothesis.

Common errors:

- Equating the nominal spot size with single-cell resolution.
- Treating an inferred cell type as observed ground truth.
- Inferring physical signalling from ligand and receptor transcripts alone.
- Generalising from one striking tissue section.

Connections: spatial transcriptomics (Section 7.6); what an expression change means (Section 7.5); protein interactions and networks (Section 11.3).

7.12.9 Statistical evidence and biological importance

A transcript has an estimated treatment log fold change of 0.35 and a nominal p-value of 0.003. Interpret these quantities without overstating the evidence.

Extensions, solution outline and common errors

Extensions:
Define the null model and tail event. Contrast a frequentist confidence interval with a Bayesian credible interval. Explain how low power, pseudoreplication, and selective reporting change the interpretation. State what additional evidence is required before describing the transcript as biologically important.
Solution outline:
The p-value is the probability, under the null model and its assumptions, of a test statistic at least as extreme as the observed statistic. It is neither the probability that the null is true nor the probability that the finding arose by chance. The coefficient estimates a conditional contrast on the model scale; its biological magnitude must be evaluated using an uncertainty interval, assay characteristics, covariate structure, and an appropriate multiplicity procedure. A 95% frequentist confidence procedure has 95% repeated-sampling coverage under the model, whereas a Bayesian credible interval has a posterior-probability interpretation conditional on the likelihood and prior. Independent replication and functional evidence are needed for a mechanistic claim.
Common errors:
Interpreting $p = 0.003$ as a 0.3% probability that the null is true; equating statistical significance with biological relevance; treating non-significance as proof of no effect; counting cells or technical replicates as independent organisms.

Connections: Section 7.4.1 (hypothesis tests and p-values); Section 7.4 (differential expression); Section 8.4 (error control in proteomics); Section 12.8 (prediction and causation).

7.12.10 Multiple testing and false-discovery control

A study tests 20,000 genes. Derive the Benjamini–Hochberg procedure and explain precisely what its false-discovery-rate guarantee does and does not mean.

Extensions, solution outline and common errors

Extensions:
Contrast FDR with family-wise error rate. Discuss independence, positive dependence, and arbitrary dependence. Explain adjusted p-values, q-values, independent filtering, and the need to define a hypothesis family before inspecting results.
Solution outline:
Let V be false rejections and R total rejections. FDR is $E[V/\max(R,1)]$, whereas FWER is $P(V \geq 1)$. Sort the m values so that $p_{(1)} \leq \cdots \leq p_{(m)}$, find the largest k satisfying $p_{(k)} \leq kq/m$, and reject through k. The original BH result assumes independent tests, with extensions to specified positive-dependence structures; Benjamini–Yekutieli provides a more conservative option under arbitrary dependence. The guarantee concerns a repeated-study expectation for a pre-specified family, not the probability that a particular reported gene is false.
Common errors:
Declaring that exactly 5% of reported genes are false; applying BH only after favourable tests have been selected; assuming correlated molecular features either invalidate every FDR method or remove the need for correction.

Connections: Section 7.4 (multiple testing); Section 8.4 (target–decoy FDR); Section 10.4 (function annotation).

7.12.11 Tumour subtypes and a prognostic signature

A group profiles 80 breast tumours by RNA-seq. Hierarchical clustering of the tumours gives three clear groups in a heat map. The group then selects the 50 genes most associated with survival, trains a classifier on them and reports a cross-validated accuracy of 95%. Evaluate both claims.

Extensions, solution outline and common errors

Extensions:

- Which distance measure, scaling and linkage would you choose, and how do they change the result?
- How would you decide whether there are three groups rather than two or five?
- How should the cross-validation have been done?
- What would convince you that the signature is clinically useful?

Solution outline:

1. Clustering always returns groups; the heat map alone is not evidence of subtypes.
2. State the choices: variance-stabilised data, the most variable genes, correlation or Euclidean distance on standardised genes, and the linkage; check that the groups survive other reasonable choices.
3. Assess the number of clusters with silhouettes or the gap statistic, and stability by resampling tumours and genes (consensus clustering).
4. Check that the groups do not follow batch, processing date or tumour purity, and validate them externally: known subtypes, enrichment of annotation, and replication in an independent cohort.
5. With $p \gg n$, selecting the 50 genes on all 80 tumours before cross-validation is selection bias; the selection must be repeated inside each fold, with nested cross-validation for any tuning.
6. Validate the final model on an independent cohort and show that it adds to clinical predictors such as grade, size and nodal status; compare with random gene sets of the same size.

Common errors:

- Treating a dendrogram as proof of discrete subtypes.
- Choosing k because the heat map "looks like" three groups.
- Reporting training or leaky cross-validated accuracy as performance on new patients.
- Interpreting the 50 selected genes as the causal drivers of outcome.

Connections: clustering and its validation, and classification with many genes (Section 7.4); landmark microarray studies (Section 7.2); regularisation, cross-validation and leakage (4).

8 Proteomics

> **From the exam description**
>
> Different proteomics methods. How are these used, algorithms.

8.1 Overview

Proteomics is the large-scale study of proteins: their identity, abundance, modifications, interactions, structures and locations. It is a broad field, built mainly on mass spectrometry, with its own instruments, algorithms and statistics, and with strong connections to the other chapters of this book. This chapter tries to cover as much of it as possible:

- mass spectrometry: history, instruments, bottom-up and top-down proteomics, fragmentation, acquisition strategies (DDA, DIA, targeted) and quantification (Section 8.2);
- algorithms for peptide identification: database search, spectral libraries, de novo sequencing with deep learning, and the prediction of spectra and retention times (Section 8.3);
- statistics: the target–decoy approach, Percolator, error control at different levels, the protein inference problem and quantitative statistics (Section 8.4);
- beyond abundance: post-translational modifications, interaction proteomics and hu.MAP, structural MS and spatial proteomics (Section 8.5);
- proteomics without MS: antibody-based proteomics and the Human Protein Atlas, affinity proteomics, multiplexed imaging and protein sequencing (Section 8.6);
- repositories, standards and the Human Proteome Project (Section 8.7).

8.2 Mass spectrometry methods

Proteomics studies the proteins themselves: which ones are present, how abundant they are, how they are modified, where they are and whom they interact with. mRNA levels predict protein levels only partly (Section 7.5), and many regulatory events, such as phosphorylation, cleavage and complex formation, happen only at the protein level. Proteomics is a broad field. Most of it is based on **mass spectrometry** (MS) (Aebersold and Mann 2003, 2016), and this section describes the instruments and experimental strategies. Methods that do not use MS are described in Section 8.6.

8.2.1 History

Mass spectrometry measures the mass-to-charge ratio (m/z) of ions. It was developed in the early twentieth century for atoms and small molecules, but proteins could not be analysed, because they could not be brought into the gas phase as intact ions. Proteomics before MS relied on **two-dimensional gel electrophoresis**, introduced by Patrick O'Farrell in 1975, which separates thousands of proteins by isoelectric point and size (O'Farrell 1975). Spots could be cut out and identified by Edman degradation, slowly, one protein at a time.

Two "soft" ionisation methods, published in 1988–1989, changed this:

- **Electrospray ionisation** (ESI): a solution is sprayed from a fine needle at high voltage, the droplets evaporate, and multiply charged ions of large molecules are released into the gas phase (Fenn et al. 1989). ESI works directly on the output of liquid chromatography.
- **Matrix-assisted laser desorption/ionisation** (MALDI): the sample is mixed with a matrix that absorbs laser light and releases mostly singly charged ions (Karas and Hillenkamp 1988; Tanaka et al. 1988).

John Fenn and Koichi Tanaka shared the 2002 Nobel Prize in Chemistry for these methods. In 1993 several groups showed that proteins could be identified by digesting them, measuring the masses of the peptides with MS (MALDI gives mostly singly charged ions, so the m/z gives the mass almost directly) and comparing these with the masses predicted from a sequence database, **peptide mass fingerprinting** (Henzel et al. 1993). Combined with the first complete genomes, which provided the databases, this started modern proteomics. The term *proteome* was coined in 1994.

8.2.2 Instruments

A mass spectrometer has an ion source, one or more **mass analysers**, which separate ions by m/z, and a detector. The main analysers differ in resolution, mass accuracy, speed and sensitivity:

- **Quadrupoles** filter ions with oscillating electric fields, letting only a chosen m/z through. They are robust and fast but have low resolution, and are used as filters and in targeted experiments.
- **Ion traps** store ions and eject them by m/z.
- **Time-of-flight** (TOF) analysers accelerate ions and measure how long they take to reach the detector; lighter ions arrive first. TOF is fast and has no upper mass limit. MALDI-TOF instruments are used for peptide mass fingerprinting and, in clinics, to identify bacteria.
- **Fourier-transform** analysers measure the frequency at which trapped ions oscillate, which depends on m/z. FT-ICR instruments use a strong magnet and give the highest resolution. The **Orbitrap**, invented by Alexander Makarov, traps ions in an electrostatic field around a central electrode and achieves very high resolution and mass accuracy (a few parts per million) without a magnet (Makarov 2000; Hu et al. 2005). Orbitrap instruments have dominated proteomics since the late 2000s.

Hybrid instruments combine analysers: a quadrupole selects precursor ions, which are fragmented in a collision cell and analysed in a TOF (**Q-TOF**) or an Orbitrap. Recent instruments add **ion mobility**, which separates ions by how fast they drift through a gas in an electric field. This depends on their charge and their size and shape (the *collisional cross section*), and adds a separation dimension between LC and m/z that can separate co-eluting peptides of similar m/z. In the timsTOF, ions are accumulated in a **trapped ion mobility spectrometry** (TIMS) device and released in order of mobility. The quadrupole is switched in step with this release, so that several precursors are selected for fragmentation within one mobility scan, which multiplies the rate of MS2 spectra (*parallel accumulation–serial fragmentation*, PASEF) (Meier et al. 2018). The same principle is used with DIA windows (dia-PASEF) (Meier et al. 2020).

The instrument never measures mass directly, only m/z. ESI produces peptide ions with several charges: a peptide of neutral mass M that has taken up z protons appears at $m/z = (M + z \cdot 1.007)/z$, where 1.007 Da is the mass of a proton. Each ion also appears as an *isotope envelope*, a series of peaks separated by $1/z$ caused by the natural abundance of ^{13}C. High resolution matters because resolving the envelope gives the charge state, and only then can the neutral monoisotopic mass be calculated from the measured m/z (*charge-state deconvolution*). This mass is what constrains the possible peptides. When this book speaks of the "mass" of a precursor or a peptide, it means the neutral mass inferred in this way.

8.2.3 Bottom-up and top-down proteomics

In **bottom-up** (shotgun) proteomics, proteins are digested into peptides, usually with trypsin, which cuts after lysine and arginine. The peptide mixture is separated by reversed-phase **liquid chromatography** (LC) coupled online to the mass spectrometer. As peptides elute, the instrument records a survey scan of intact peptide ions (**MS1**), selects precursor ions, fragments them and records the fragment spectra (**MS2**). This is **tandem mass spectrometry** (LC-MS/MS). Peptides are easier to separate, ionise and fragment than whole proteins, but the connection between a peptide and the protein it came from is lost, which leads to the protein inference problem (Section 8.4).

In **top-down** proteomics, intact proteins are measured and fragmented. This preserves the full *proteoform*, the exact combination of sequence variants, splice isoform and modifications of each molecule, but it is technically much harder, especially for large proteins (Toby et al. 2016). Large-scale top-down studies have shown that the products of one gene often appear as several distinct intact proteoforms (Tran et al. 2011). Top-down is not simply a better form of bottom-up proteomics. It trades depth and throughput for molecular resolution, and separation, charge-state deconvolution and fragmentation all become harder as the mass of the protein grows.

The difference matters for interpretation. Bottom-up depth came from digestion combined with multidimensional chromatography, as in the MudPIT analysis of the yeast proteome (Washburn et al. 2001), but digestion breaks the link between modifications on different peptides. Observing a phosphorylated peptide and an acetylated peptide from the same gene does not show that both modifications occur on the same molecule. **Middle-down** proteomics, which analyses longer peptides produced by less frequently cutting proteases, lies in between: it keeps more information on which modifications co-occur, for example on histone tails, while remaining easier than the analysis of intact proteins. A tryptic peptide can support a gene-level protein group, but it rarely establishes the intact proteoform.

8.2.4 Fragmentation and b- and y-ions

The most common fragmentation method, **collision-induced dissociation** (CID), and its beam-type variant HCD, collide peptide ions with gas molecules. The peptide breaks mainly at the amide bonds of the backbone. If the charge stays on the N-terminal fragment it is a **b-ion**; if it stays on the C-terminal fragment it is a **y-ion**. A complete series of b-ions (or y-ions) forms a ladder in which the mass difference between consecutive peaks equals the mass of one amino acid residue, so the sequence can in principle be read from the spectrum. In practice spectra are incomplete and noisy, and contain other ion types, neutral losses and peaks from co-fragmented peptides. Leucine and isoleucine have the same mass and cannot be distinguished. Electron-based fragmentation (ETD and ECD) cleaves a different backbone bond, producing c- and z-ions, and preserves labile modifications such as phosphorylation and glycosylation.

8.2.5 Acquisition strategies

How the instrument decides which ions to fragment determines what is measured:

- **Data-dependent acquisition** (DDA): after each MS1 scan, the instrument selects the most intense precursors (the *top N*) for fragmentation, and excludes them for a while so that other peptides get a chance. Each MS2 spectrum ideally comes from one peptide, which makes identification simple. But selection is semi-random, so different runs identify different peptides, and low-abundance peptides are missed, which causes many **missing values** across samples.

- **Data-independent acquisition** (DIA), introduced as **SWATH-MS** (Gillet et al. 2012; Ludwig et al. 2018), fragments *all* ions in a series of wide m/z windows that together cover the mass range, repeated throughout the LC run. Every peptide is fragmented in every run, which gives more complete and reproducible quantification. The price is that each MS2 spectrum is a mixture of fragments from many peptides, which requires dedicated algorithms (Section 8.3).
- **Targeted acquisition**: in **selected reaction monitoring** (SRM, also called MRM) on triple-quadrupole instruments, the instrument monitors only predefined pairs of precursor and fragment masses for a list of peptides (Lange et al. 2008). **Parallel reaction monitoring** (PRM) records full fragment spectra of selected precursors on high-resolution instruments (Peterson et al. 2012). Targeted methods are the most sensitive and accurate, and are used to validate biomarkers, but they measure only a few hundred peptides chosen in advance. A multi-laboratory study showed that targeted assays with isotope-labelled internal standards can be transferred between laboratories with good reproducibility, provided that sample preparation is standardised (Addona et al. 2009).

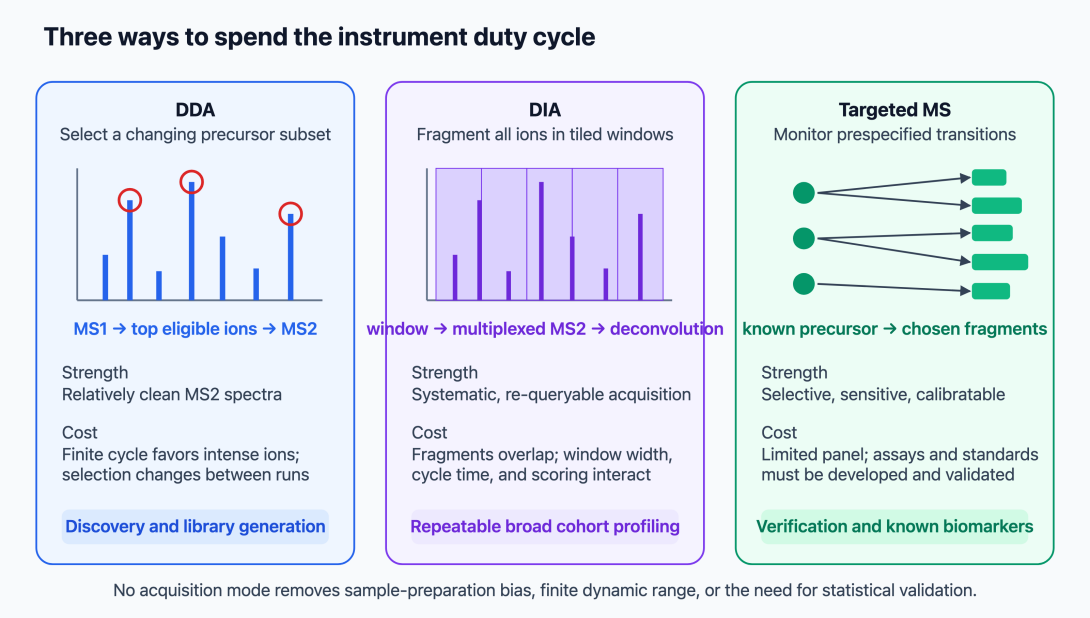


Figure 8.1: The three acquisition strategies. Data-dependent acquisition fragments a changing subset of the most intense precursors, data-independent acquisition fragments all ions in a series of m/z windows, and targeted acquisition monitors predefined precursor–fragment transitions. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

All three strategies divide a limited **duty cycle** between scans. Chromatographic peaks last only seconds, and each peptide must be sampled several times across its peak for its area to be estimated. In DIA, the width of the isolation windows sets the trade-off: narrow windows mix fewer precursors and reduce interference, but more windows are needed to cover the mass range, which lengthens the cycle. Wide windows are fast but give more complex spectra. Variable-width windows, narrower where precursors are dense, are a common compromise. Offline fractionation before LC-MS reduces the complexity of each injection and increases depth, at the cost of instrument time and a more complicated pattern of missing values.

8.2.6 Quantification

Mass spectrometry is not inherently quantitative, since different peptides ionise with very different efficiency. Quantification therefore compares the *same* peptide between samples:

- **Label-free quantification** compares peptide intensities (the area of the MS1 peak over the elution time) or, more crudely, the number of spectra identified for each protein (*spectral counting*) between separate runs. It requires careful alignment of retention times and normalisation between runs. The MaxLFQ algorithm uses ratios of shared peptides between pairs of samples (Cox et al. 2014).
- **Metabolic labelling**: in **SILAC**, cells are grown with amino acids (usually arginine and lysine) containing heavy isotopes (Ong et al. 2002). Light and heavy samples are mixed before processing, and each peptide appears as a pair of peaks whose intensity ratio gives the relative abundance. Mixing early minimises technical variation.
- **Isobaric chemical tags**, **iTRAQ** (Ross et al. 2004) and **TMT** (Thompson et al. 2003), label peptides from up to 18 or more samples with tags of the same total mass, so that the same peptide from all samples appears as a single MS1 peak. On fragmentation, each tag releases a *reporter ion* of a different mass, whose intensities give the relative abundance in each sample. Multiplexing reduces missing values, but co-isolated peptides contribute reporter ions too, which compresses the measured ratios (*ratio compression*).
- **Absolute quantification** uses spiked-in heavy synthetic peptides of known amount (Gerber et al. 2003), or estimates absolute abundance from the sum of peptide intensities divided by the number of possible peptides (iBAQ) (Schwanhäusser et al. 2011).

None of these values is a count of molecules. Each method has its own biases. Label-free intensities depend on how well each peptide ionises, which is why only the same peptide can be compared between samples. Spectral counts are coarse, especially for proteins with few spectra, and favour long proteins, which give more peptides. Reporter-ion ratios are compressed by co-isolated peptides. Protein-level abundances are therefore relative measures between samples, unless they are calibrated with standards of known amount, as in AQUA. iBAQ values are approximations that are useful for ranking proteins within a sample, not exact copy numbers.

8.2.7 Experimental design and quality control

A proteomics result depends on a long chain of steps, from sampling and digestion to peptide identification, protein inference and statistics (Figure 8.2). Each step adds its own errors, and later computation cannot recover proteins that were lost or degraded before measurement.

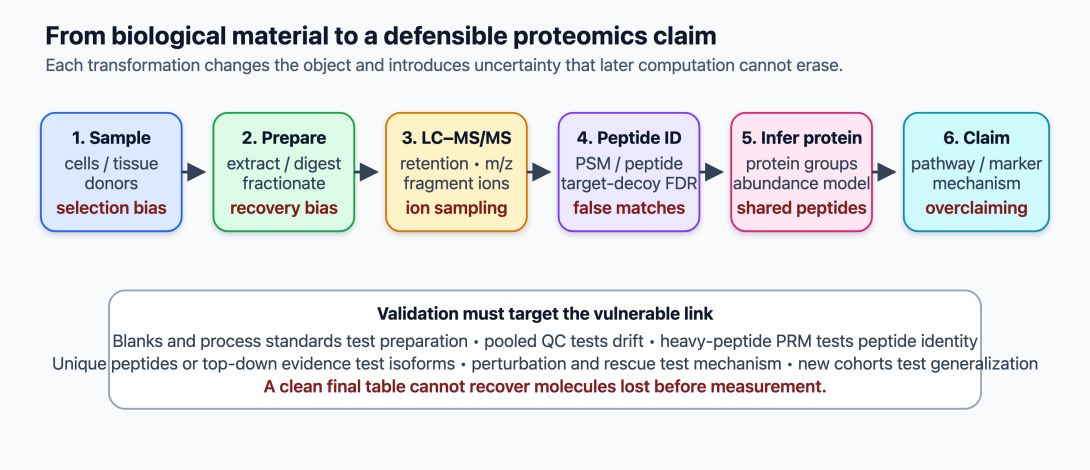


Figure 8.2: The chain of inference in a proteomics experiment. Each step, from sampling and preparation through LC-MS/MS, peptide identification and protein inference, introduces its own errors, and validation should target the weakest link. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

The quantification strategies differ in how they handle the design:

- **SILAC** samples are mixed early, so later handling affects them equally. Incomplete labelling and the metabolic conversion of labelled arginine to proline can bias ratios, and a *label swap*, in which the heavy and light labels are exchanged between conditions in a replicate, reveals such bias. SILAC is limited to cells that can be grown in culture.
- **TMT** experiments with more samples than channels are split into several *plexes*. Comparisons within a plex are the most precise. A pooled reference sample included in every plex allows them to be put on a common scale (*bridging*), and conditions should be distributed evenly and randomly across plexes and channels. If a condition is confounded with a plex, no normalisation can separate the two. Ratio compression from co-isolated peptides can be reduced by narrower isolation windows or by a further round of fragmentation before reporter-ion measurement (MS3) (Ting et al. 2011).
- **Label-free** experiments measure every sample in a separate run, so drift of the LC and the instrument over time, and the order of the runs, must be monitored and randomised with respect to the conditions.

Quality control should be able to locate a failure, not only detect it. Common controls are blank injections, which reveal *carryover* from the previous sample; a pooled QC sample injected repeatedly during the run, which monitors drift; and standards added at different stages. A peptide standard added just before injection tests only the LC-MS system, whereas a protein standard added before digestion also tests digestion and recovery. Useful summaries include the rate of missed cleavages, mass errors, retention-time stability, the number of identifications per run and the coefficient of variation of the pooled QC samples. Finally, the unit of replication is the independent biological sample: repeated injections of the same digest, or several TMT channels made from one sample, are technical replicates.

See also: further explanation in Section D.8.1; related Wikipedia articles in Section E.8.1.

Sources

Parts of this section are adapted from the Wikipedia articles Proteomics, Mass spectrometry, Electrospray ionization, Matrix-assisted laser desorption/ionization, Orbitrap, Tandem mass spectrometry, Top-down proteomics, Data-independent acquisition, Stable isotope labeling by amino acids in cell culture and Isobaric labeling (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Proteomics and Mass-Spectrometric Inference” (CC BY 4.0), used with permission.

8.3 Algorithms for peptide identification

A single LC-MS/MS run produces tens of thousands of fragment spectra. The central computational task is to work out which peptide produced each spectrum, a **peptide–spectrum match** (PSM). There are three approaches: comparing spectra with sequences in a database, comparing them with previously observed spectra, or reading the sequence directly from the spectrum (*de novo* sequencing). Since around 2019, deep learning has improved all three (Nesvizhskii 2010).

Terminology: from spectrum to protein group

- A **precursor** is the ion selected from the MS1 scan for fragmentation. It is a **peptide ion**: one peptide with a particular charge state (and particular modifications). The same peptide can give several precursors, for example with charge 2+ and 3+.

- A **peptide sequence** is the amino acid sequence itself, independent of charge. The same sequence with different modifications is sometimes called a different *peptidoform*.
- A **PSM** assigns one fragment spectrum to one peptide ion. Many PSMs can support the same peptide sequence.
- A **protein group** is the set of proteins that the observed peptides cannot distinguish (Section 8.4). It is usually what is reported, rather than a single protein or isoform.

Counts, error rates and quantities can refer to any of these levels, and a statement such as "5,000 identifications at 1% FDR" should say which.

8.3.1 Database search

Database search is the standard approach for organisms with a sequenced genome. For each experimental spectrum:

1. The protein database is digested *in silico* with the rules of the protease, and **candidate peptides** whose mass matches the precursor mass within a tolerance (a few ppm on modern instruments) are selected. The precursor mass is not measured directly: it is calculated from the measured m/z and the charge state inferred from the isotope envelope (Section 8.2).
2. For each candidate, a **theoretical spectrum** is generated, with the expected b- and y-ions.
3. A **score** measures the agreement between the experimental and theoretical spectra, and the best-scoring candidate is reported.

SEQUEST, from John Yates's laboratory, was the first such program (Eng et al. 1994). Its main score, *XCorr*, is the cross-correlation between the experimental spectrum and the theoretical one, minus the average correlation at shifted offsets, which corrects for random agreement. **Mascot** introduced a probability-based score, the probability that the observed number of matching peaks arises by chance (Perkins et al. 1999). Many others followed: X!Tandem (Craig and Beavis 2004), Andromeda, integrated in the widely used MaxQuant platform (Cox and Mann 2008; Cox et al. 2011), and MS-GF+, which computes exact score distributions with dynamic programming (Kim and Pevzner 2014).

A search score ranks the candidate peptides for one spectrum. A high score is not by itself a calibrated probability that the match is correct: score distributions differ between spectra, charge states and search spaces, and even a probability-based score such as Mascot's describes how unlikely the match is under a random-match model for that spectrum, not the error rate of a list of accepted PSMs. The FDR of a set of identifications must be estimated separately, with the target–decoy approach or another explicit error model (Section 8.4).

The size of the **search space** is a central issue. Allowing for variable post-translational modifications, non-specific cleavage, or sequence variants multiplies the number of candidates, which increases both search time and the chance that a random peptide scores well. *Open search* allows any mass difference between precursor and peptide, so that unexpected modifications can be discovered; MSFragger made this fast with an indexed fragment-ion database (Kong et al. 2017). In **proteogenomics**, the database is extended with peptides translated from genomic variants, new open reading frames or RNA-seq transcripts, which can reveal unannotated genes and improve gene annotation (Section 5.6). Such databases can be specific to the sample, for example built from the genome or RNA-seq data of the same tumour, so that peptides carrying the patient's own variants can be detected. Novel peptides need particularly strict error control: the extended database is larger, and a non-canonical peptide is a priori less likely to be correct than a canonical one, so its FDR should be estimated separately from that of the known peptides.

8.3.2 Spectral library search

Peptides that have been identified before have characteristic spectra, with reproducible relative intensities that the simple theoretical spectra of database search ignore. **Spectral library search** compares each spectrum with a library of previously identified, high-quality spectra, for example with a normalised dot product (Lam et al. 2007). It is faster and more sensitive than database search, but it can only identify peptides already in the library. Spectral libraries are also the basis of the analysis of DIA data.

8.3.3 De novo sequencing

De novo sequencing infers the peptide sequence from the spectrum alone, without a database. It is needed for organisms without a sequenced genome, for antibodies, whose sequences are generated by somatic recombination and mutation, for immunopeptidomics (peptides presented by MHC molecules, which are not tryptic), and for finding unexpected sequences.

The classical algorithms represent the spectrum as a **spectrum graph**: each peak is a node, interpreted as a possible b- or y-ion, and two nodes are connected by an edge if their mass difference equals the mass of one (or two) amino acids. The best path through the graph, found with dynamic programming, is the most likely sequence (Dančík et al. 1999). PEAKS combined this with careful scoring and became the standard commercial tool (Ma et al. 2003). The difficulty is that real spectra have missing fragment ions, noise peaks and ambiguous interpretations, so de novo methods were much less accurate than database search.

Deep learning has changed this. DeepNovo combined convolutional and recurrent neural networks to predict the sequence amino acid by amino acid (Tran et al. 2017). **Casanovo** treats de novo sequencing as a translation problem, from a "sentence" of peaks to a sequence of amino acids, using a transformer model, the architecture of large language models (Yilmaz et al. 2024). **InstaNovo** combines a transformer with a diffusion model that iteratively refines the predicted sequence, and was trained on tens of millions of spectra (Eloff et al. 2025). Such models, trained on very large collections of identified spectra from public repositories, are sometimes described as **foundation models** for peptide sequencing. They now identify many peptides with an accuracy approaching database search, and can find peptides that database search misses, although error control for de novo results is still less mature (Section 8.4).

8.3.4 Predicting spectra and retention times

The behaviour of a peptide in an LC-MS/MS experiment is largely determined by its sequence: its retention time on the chromatography column, the relative intensities of its fragment ions and, in ion-mobility instruments, its collisional cross section. Deep learning models now predict these properties accurately. **Prosit** predicts fragment intensities and retention times with a recurrent neural network trained on spectra of hundreds of thousands of synthetic peptides (Gessulat et al. 2019). **MS²PIP** predicts fragment intensities with gradient-boosted trees and deep learning (Gabriels et al. 2019). **AlphaPeptDeep** is a modular framework of neural network models, including transformers, for predicting spectra, retention times and collisional cross sections, which can easily be retrained for new modifications or instruments (Zeng et al. 2022).

Predicted properties are used in three ways:

- **Rescoring**: the agreement between the observed spectrum and the predicted one, which is much more specific than a simple theoretical spectrum, is added as a feature in rescoring (Section 8.4). This gives many more identifications at the same FDR, particularly for difficult cases such as non-tryptic immunopeptides.

- **Predicted spectral libraries**: libraries can be predicted for every peptide in a proteome, removing the need to build them experimentally.
- **DIA analysis**: DIA spectra mix fragments from many peptides, so they are analysed in a *peptide-centric* way. For each peptide in a (measured or predicted) library, the software extracts the chromatograms of its expected fragment ions at its expected retention time, and scores how well they co-elute and match the expected intensities, as in OpenSWATH (Röst et al. 2014). DIA-NN uses neural networks to separate true from false peptide signals and to correct for interference (Demichev et al. 2019). With predicted libraries, DIA can be analysed without any experimentally built library ("library-free" DIA).

See also: further explanation in Section D.8.2; related Wikipedia articles in Section E.8.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Peptide mass fingerprinting, Tandem mass spectrometry, De novo peptide sequencing, Proteogenomics and Shotgun proteomics (CC BY-SA 4.0, retrieved 4 October 2026).

8.4 Statistics and error control

Every spectrum gets a best-matching peptide, whether or not the correct peptide is in the database, and many spectra are noise or come from peptides that are not considered, so a large fraction of the top-scoring PSMs in a typical experiment are wrong. Error control is therefore the core statistical problem of proteomics, and it is addressed with ideas closely related to the false discovery rate in expression analysis (Section 7.4) (Nesvizhskii 2010).

8.4.1 The target–decoy approach

The **target–decoy** strategy estimates the number of false matches empirically (Elias and Gygi 2007). The spectra are searched against a database that contains the real (*target*) protein sequences and an equal number of *decoy* sequences that cannot be present in the sample, typically the target sequences reversed or shuffled. Decoys have the same size and amino acid composition as the targets, so a wrong match is equally likely to hit a target or a decoy. Above any score threshold, the number of decoy matches therefore estimates the number of false target matches, and

$$\widehat{\mathrm{FDR}} = \frac{\#\text{decoy matches above threshold}}{\#\text{target matches above threshold}}.$$

The threshold is chosen to give the desired FDR, usually 1% (here at the level of PSMs; see below for other levels), and, as in expression analysis, a **q-value** is computed for each PSM as the lowest FDR at which it would be accepted. The approach makes no assumptions about the score distribution and works with any search engine, which is why it became the standard. It assumes that decoys behave like random false matches, which can fail, for example if the decoy construction preserves too much of the target sequences or if the search space is very large.

In the usual *concatenated* search, targets and decoys compete for each spectrum, and only the best match is kept. Implementations differ in details, for example estimating the FDR as $(D+1)/T$ for D decoy and T target matches, which is slightly conservative, or rescaling when the decoy database is not the same size as the target database. With larger scores better, the q-value of a PSM with score s_i is

$$q_i = \min_{t \le s_i} \widehat{\mathrm{FDR}}(t),$$

the lowest estimated FDR of any threshold at which it is accepted. Two points are often misunderstood. First, an FDR is the expected proportion of errors in an accepted *list*: a PSM in a list at 1% FDR does not have a 1% chance of being wrong, since high-scoring PSMs are almost always right and those near the threshold much more often wrong. Second, a global FDR can hide much higher error rates in subgroups, such as modified peptides, unusual charge states, proteins identified by a single peptide or novel peptides (Section 8.3), which may need to be controlled separately.

Earlier, and complementary, approaches model the score distribution directly. **PeptideProphet** fits a mixture of two distributions, one for correct and one for incorrect PSMs, with the EM algorithm, and gives each PSM a **posterior error probability** (PEP), the probability that this particular match is wrong (Keller et al. 2002). The PEP is a local measure, like Efron's local FDR, while the q-value is a property of the whole list above a threshold.

8.4.2 Rescoring with machine learning: Percolator

A search engine score is only one of many features that distinguish correct from incorrect PSMs. Others include the mass error, the difference from the second-best score, the number of missed cleavages, the charge state, the retention time and, today, the similarity to a predicted spectrum (Section 8.3). **Percolator** combines them with a **semi-supervised** machine learning method (Käll et al. 2007). Decoy PSMs are known negatives. The best-scoring target PSMs are taken as provisional positives. A support vector machine is trained to separate the two sets, all PSMs are rescored, and the procedure is repeated with the new positives. Cross-validation prevents the classifier from overfitting. Percolator typically increases the number of PSMs or peptides accepted at 1% FDR by tens of percent, and the same principle is used in most current pipelines.

8.4.3 Levels of error control

Error rates can be controlled at several levels, and they are not the same:

- **PSM level**: each spectrum–peptide match.
- **Peptide level**: unique peptide sequences. Several spectra can match the same peptide, and correct peptides tend to be identified by more spectra than incorrect ones.
- **Protein level**: a protein (or protein group) is reported if at least one of its peptides is. Since false peptides are spread randomly over all proteins, while true peptides concentrate on a smaller number of proteins, a 1% FDR at the peptide level can correspond to a much higher FDR at the protein level, especially in large data sets. Protein-level FDR must therefore be estimated separately, for example with the *picked* target–decoy approach, which treats each target protein and its decoy as a pair (Savitski et al. 2015).

A "1% FDR" is therefore incomplete without its level: 1% of PSMs, of peptides or of protein groups are different statements, and an FDR controlled at one level does not control it at another. At the protein level, protein inference (below) adds a further ambiguity, because what is counted depends on how peptides are assigned to proteins and groups.

8.4.4 The protein inference problem

Bottom-up proteomics identifies peptides, but the biological questions are about proteins. This step is not a technical detail: every protein list, protein-level FDR and protein quantity in a proteomics paper depends on it. Many peptides are shared between several proteins: paralogues, isoforms produced by alternative splicing, and proteins with shared domains. Deciding which proteins are actually present is the **protein inference problem**. Observed peptides therefore rarely "identify the protein" directly. They are evidence for a set of candidate proteins, and the data often support only a group of them (Figure 8.3). The problem is most severe for splice isoforms, paralogues, immunoglobulins and other families of highly similar proteins (Nesvizhskii and Aebersold 2005).

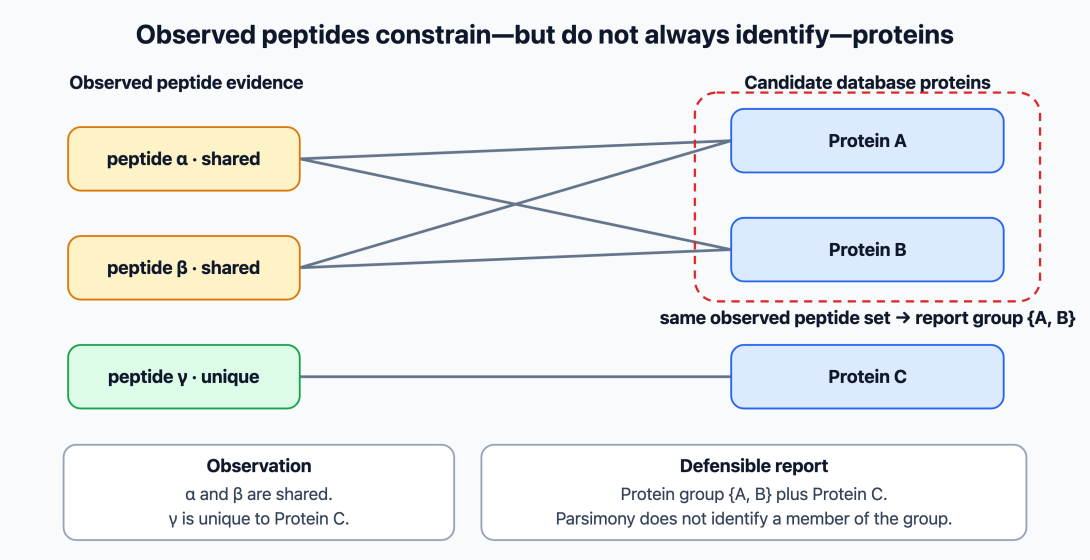


Figure 8.3: Protein inference. Peptides α and β are shared by proteins A and B, so they support only the group {A, B}, not either member alone, while peptide γ is unique to protein C. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

The common solution follows the principle of **parsimony** (Occam's razor): report the smallest set of proteins that explains all observed peptides. Proteins that cannot be distinguished by the observed peptides are reported together as a **protein group**. ProteinProphet introduced a probabilistic version, which apportions shared peptides between proteins according to the evidence from their unique peptides (Nesvizhskii et al. 2003). Parsimony is a heuristic: a protein may be present even if all its peptides are shared with another, and conversely. Isoforms, in particular, can rarely be distinguished, because the peptides that would discriminate between them are often not observed. Software also differs in how shared peptides are used for quantification: some use only peptides unique to a group, others assign each shared peptide to the group with the most other evidence (the *razor* peptides of MaxQuant (Cox and Mann 2008)). The same peptide data can thus give different protein lists and quantities, depending on the rules used. Quantification makes this worse: if a shared peptide changes between conditions but the peptides unique to each isoform do not, the change cannot be assigned to one isoform. A transparent report separates unique, razor and shared peptides, and checks whether the conclusions change when shared or discordant peptides are excluded.

8.4.5 Statistics of quantification

Quantitative proteomics faces the same statistical problems as transcriptomics (normalisation, replicates, multiple testing, Section 7.4), with two additional complications. First, protein abundances are estimated from several peptides, which must be summarised robustly, since some peptides are shared or modified.

Second, **missing values** are frequent, often tens of percent of the peptide values in a label-free DDA data set, and they are not missing at random (Lazar et al. 2016). Some are *censored*: the peptide is present but below the detection limit, so it is more likely to be missing if it has low abundance. Others arise because the peptide was not selected for fragmentation or not identified in that run, which is closer to random. The two need different treatment. Simply ignoring missing values discards proteins that are present in one condition and absent from the other, which may be the largest changes, while replacing them with small values creates large fold changes that may be artefacts. Imputation with small values, models that treat low values as censored, or methods that transfer identifications between runs (*match between runs*) all affect the list of differentially abundant proteins considerably, and the chosen approach should be reported. Dedicated statistical packages, such as MSstats, fit linear mixed models at the peptide level and test proteins for differential abundance (Choi et al. 2014), and limma (Section 7.4) is also widely used.

8.4.6 Normalisation and differential abundance

A simple model of a label-free experiment helps to see what can and cannot be compared. The measured intensity of peptide p from protein i in sample s is approximately

$$I_{pis} \approx A_{is}\, R_{pi}\, E_{ps} + \varepsilon_{pis},$$

where A_{is} is the abundance of the protein, R_{pi} the response of the peptide (how well it is recovered and ionised), E_{ps} run-specific effects such as loading and interference, and ε_{pis} measurement error. Since R_{pi} varies enormously between peptides, intensities of different peptides cannot be compared directly, but the ratio for the same peptide between samples cancels it. Intensities are usually log-transformed, which makes these multiplicative effects additive.

Normalisation removes differences in E_{ps} that affect all peptides in a run, such as the amount loaded. Median or similar scaling assumes that most proteins do not change, and fails if most measured proteins really shift in one direction, or if the composition of the samples differs, for example between tissues. Normalising instead to internal standards, cell numbers or total protein answers a different question, and the choice should be stated. MaxLFQ postpones normalisation and estimates protein profiles from the ratios of peptides shared between pairs of samples (Cox et al. 2014).

For testing, the design (paired samples, batches, TMT plexes, repeated measurements) should be specified in the model before the results are examined. Moderated linear models and mixed models, as in limma and MSstats, borrow strength across proteins when replicates are few. Two different FDRs are involved in a typical study. The identification FDR decides which peptides and proteins enter the table; the FDR of the differential analysis decides which of them differ between conditions. Neither controls the other, and neither validates a subsequent pathway enrichment, which should use the measured proteins, not the whole proteome, as its background (Section 7.5).

See also: further explanation in Section D.8.3; related Wikipedia articles in Section E.8.3.

i Sources

Parts of this section are adapted from the Wikipedia articles False discovery rate, Shotgun proteomics and Semi-supervised learning (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter “Proteomics and Mass-Spectrometric Inference” (CC BY 4.0), used with permission.

8.5 Proteomics beyond protein abundance

Mass spectrometry measures much more than which proteins are present and how abundant they are. This section describes the analysis of post-translational modifications, interactions and complexes, protein structure, and the location of proteins in cells and tissues.

8.5.1 Post-translational modifications

Proteins are modified after translation by the addition of chemical groups (phosphate, acetyl, methyl, sugars), small proteins (ubiquitin, SUMO) or lipids, and by proteolytic cleavage. These **post-translational modifications** (PTMs) control activity, localisation, interactions and degradation, and they can only be measured at the protein level.

Phosphoproteomics is the most developed example. Phosphorylation of serine, threonine and tyrosine by kinases is the main mechanism of signalling in cells. Phosphopeptides are rare and their phosphorylation is often substoichiometric, so they are **enriched** before MS, typically with titanium dioxide or immobilised metal ions, or with antibodies against phosphotyrosine. Large-scale studies then measure tens of thousands of phosphorylation sites, and how they change over time after a stimulus (Olsen et al. 2006). Two computational problems are specific to PTMs:

- **Search space**: allowing each serine, threonine and tyrosine to be phosphorylated or not multiplies the number of candidate peptides.
- **Site localisation**: a peptide may contain several possible sites, and the spectrum must show which one carries the modification. Scores such as the Ascore compute the probability of each localisation from the site-determining fragment ions (Beausoleil et al. 2006). Identifying a phosphopeptide therefore does not by itself localise the phosphate to one residue: if the fragments that would distinguish neighbouring sites are missing, the site is ambiguous. Results are reported per site with a localisation probability (for example, Olsen et al. (2006) called sites with a probability of at least 0.75 "class I"), and the confidence in the site is separate from the FDR of the peptide identification.

Similar enrichment strategies exist for ubiquitination (antibodies against the di-glycine remnant left on lysines after trypsin digestion), acetylation and glycosylation. Interpreting PTM data, for example to infer which kinases are active, relies on databases of known sites and kinase substrate motifs.

8.5.2 Abundance and occupancy of modifications

A PTM data set answers up to four different questions, and the evidence for each is different:

1. Was a modified peptide identified? This rests on the database search and its FDR.
2. Which residue carries the modification? This is the localisation problem above, with its own error rate.
3. How abundant is the modified peptide? This is a relative quantity from the ion signal, like any other peptide.
4. What fraction of the molecules of the protein carry the modification (the *occupancy* or stoichiometry)? This requires the modified and the unmodified forms, and a correction for their different responses in the mass spectrometer.

Because modified peptides are enriched before measurement, an increase in a phosphopeptide can mean that the occupancy of the site rose, that the total amount of the protein rose, or that the enrichment recovered more material. Phosphoproteomics data are therefore usually interpreted together with a

measurement of the unenriched proteome, so that changes in sites can be separated from changes in protein abundance. Mechanistic claims about a single site are typically confirmed with synthetic modified peptides, targeted MS, or perturbation of the responsible kinase.

8.5.3 Interaction proteomics

Proteins act in complexes and pathways. Mass spectrometry is one of the two main experimental sources of protein interaction data, the other being the yeast two-hybrid system. The interaction networks built from these data are discussed in Chapter 11.

- **Affinity purification–mass spectrometry** (AP-MS): a tagged *bait* protein is purified from cells together with the proteins bound to it, the *preys*, which are identified by MS. The first proteome-scale studies mapped complexes in yeast (Gavin et al. 2002; Ho et al. 2002), and the BioPlex project has purified more than 10,000 human baits in two cell lines (Huttlin et al. 2021). AP-MS detects *co-complex* membership, not direct binding, and many identified preys are contaminants that bind to the beads or are simply abundant. Scoring methods, such as SAINT, compare the abundance of each prey with negative controls to compute the probability of a true interaction (Choi et al. 2010), and the CRAPome collects contaminants seen in control experiments (Mellacheruvu et al. 2013).
- **Proximity labelling** (BioID, APEX) fuses the bait to an enzyme that biotinylates all proteins within about 10 nm in the living cell, which are then purified and identified (Roux et al. 2012). It captures weak and transient interactions and works for insoluble structures, but labels neighbours as well as binding partners.
- **Co-fractionation MS** (CF-MS) separates native complexes from a cell extract by chromatography into many fractions and measures the proteins in each. Proteins that are always found in the same fractions are likely to be in the same complex. No tagging is needed, so the method can be applied to any organism or tissue (Havugimana et al. 2012; Wan et al. 2015).

8.5.3.1 hu.MAP: integrating proteomics experiments into a map of complexes

Each of these methods has its own biases and errors, so combining them gives more reliable maps. **hu.MAP**, developed by Kevin Drew and Edward Marcotte, integrates large public data sets of AP-MS (including BioPlex), CF-MS and proximity labelling. The approach is a typical machine-learning pipeline:

1. For each pair of proteins, features are computed from every experiment: co-purification scores from AP-MS, correlation of fractionation profiles from CF-MS, co-labelling from proximity labelling.
2. A classifier is trained to distinguish pairs of proteins known to be in the same complex, from curated databases, from pairs that are not.
3. The classifier scores all pairs, giving a weighted network of co-complex interactions.
4. The network is **clustered** to define complexes, at several levels of confidence.

hu.MAP 2.0 integrated more than 15,000 experiments into nearly 7,000 complexes (Drew et al. 2021), and hu.MAP 3.0 more than 25,000 experiments into more than 15,000 complexes, which place nearly 70% of human proteins in a physical context. hu.MAP 3.0 also used AlphaFold models of the complexes to identify subunits that are mutually exclusive (Fischer et al. 2025). As with any such map, the predictions are only as good as the training data, and the evaluation must avoid using the same known complexes for training and testing. hu.MAP is used as one of the large interactome maps in Chapter 11 and is described as a database in Section 2.7.

8.5.4 Structural mass spectrometry

MS also provides structural information, which complements X-ray crystallography, cryo-EM and structure prediction (Chapter 9):

- **Cross-linking MS** (XL-MS) covalently links amino acids that are close in space, typically lysines within about 30 Å, digests the proteins and identifies the cross-linked peptide pairs (O'Reilly and Rappsilber 2018). Cross-links give distance restraints for modelling complexes and reveal interactions in cells. The computational problem is hard: the search space is all pairs of peptides, which grows quadratically with the database, and the target–decoy approach must be adapted to pairs.
- **Hydrogen–deuterium exchange** (HDX-MS) measures how quickly backbone amide hydrogens exchange with deuterium from the solvent, which reports on solvent accessibility and dynamics.
- **Native MS** measures intact complexes, giving their stoichiometry.
- **Limited proteolysis** (LiP-MS) (Schopper et al. 2017) and **thermal proteome profiling** (Savitski et al. 2014) detect changes in protein structure or stability across the whole proteome, for example when a drug binds its targets.

8.5.5 Spatial proteomics

Where a protein is located is part of its function. **Organelle proteomics** separates cell compartments by density-gradient centrifugation and assigns proteins to organelles from their distribution across fractions, with machine-learning classifiers trained on known markers (Christoforou et al. 2016). Antibody-based imaging of human proteins in cells, in the Human Protein Atlas, has produced a subcellular map showing that about half of all proteins are found in more than one location (Thul et al. 2017) (Section 8.6).

In tissues, **MS imaging** with MALDI scans a laser across a tissue section and records a mass spectrum at each position (Caprioli et al. 1997). **Deep visual proteomics** combines microscopy and machine learning to identify cells of interest, cuts them out with a laser and analyses them by sensitive MS (Mund et al. 2022). **Single-cell proteomics** by MS, first demonstrated with isobaric labelling (Budnik et al. 2018), now measures a few thousand proteins in individual cells, still far fewer than single-cell transcriptomics (Section 7.6). Image-based methods for protein localisation and for tissues are described in Section 13.5 and Section 13.6.

See also: further explanation in Section D.8.4; related Wikipedia articles in Section E.8.4.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Post-translational modification, Phosphoproteomics, Protein–protein interaction, Hydrogen–deuterium exchange and Mass spectrometry imaging (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Proteomics and Mass-Spectrometric Inference" (CC BY 4.0), used with permission.

8.6 Proteomics without mass spectrometry

Not all proteomics uses mass spectrometry. Methods based on binding reagents, such as antibodies and aptamers, are more sensitive for some applications, can be applied to thousands of samples, and can show where proteins are located in tissues and cells.

8.6.1 Antibody-based proteomics and the Human Protein Atlas

Antibodies have long been used to detect individual proteins: in Western blots, in enzyme-linked immunosorbent assays (ELISA), and in immunohistochemistry and immunofluorescence, which show where a protein is in a tissue or a cell. The **Human Protein Atlas** (HPA), started in Sweden in 2003 and led by Mathias Uhlén, has scaled this up to the whole human proteome. The project generated antibodies against most human proteins and used them to stain tissues and cell lines, combined with transcriptomics (Uhlén et al. 2015). It is organised as several atlases:

- the **tissue atlas**, describing the expression and localisation of proteins across human tissues and organs, and classifying genes by their tissue specificity (Uhlén et al. 2015);
- the **subcellular atlas**, based on immunofluorescence of cell lines, showing the location of proteins in organelles and substructures (Thul et al. 2017);
- the **pathology atlas**, linking expression to survival in cancers (Uhlen et al. 2017);
- the **single-cell type atlas**, based on single-cell RNA-seq (Karlsson et al. 2021), and further atlases for blood, brain and other systems.

The reliability of antibody-based data depends on the specificity of the antibodies, and many commercial antibodies bind other proteins than their target. The HPA consortium and others therefore proposed standard strategies for **antibody validation**, such as genetic knockdown, comparison with an independent antibody, or correlation with RNA levels (Uhlen et al. 2016). Even a validated antibody gives weaker evidence than MS: a signal shows that something the antibody binds is present, but, unlike a peptide identified by MS, it does not report the sequence of the molecule that was detected. Antibody detection of a protein should therefore not be treated as equivalent to sequence-resolved identification. The HPA data are freely available and are described as a database in Section 2.8.

8.6.2 Affinity proteomics of body fluids

Blood plasma is the most accessible human sample, but its protein concentrations span more than ten orders of magnitude, which makes deep measurement by MS difficult. **Affinity proteomics** measures predefined panels of proteins with binding reagents:

- The **proximity extension assay** (PEA), developed by Olink in Uppsala, uses pairs of antibodies coupled to DNA oligonucleotides. When both antibodies bind the same protein, the oligonucleotides come close together, hybridise and are extended by a polymerase into a DNA barcode, which is then quantified by PCR or sequencing (Assarsson et al. 2014). Only the specific pair produces a signal, which reduces cross-reactivity.
- **SomaScan** uses *aptamers*, short single-stranded DNA molecules selected to bind specific proteins, and measures them on DNA arrays (Gold et al. 2010).

Both platforms now measure thousands of proteins in microlitres of plasma, and have been applied to tens of thousands of individuals. Combined with genotypes, such studies map *protein quantitative trait loci* (pQTLs) and link plasma proteins to disease, as in the UK Biobank Pharma Proteomics Project (Sun et al. 2023). Affinity assays measure binding to an epitope, not the protein sequence, so specificity remains the main source of uncertainty. The two platforms agree well for some proteins and poorly for others, and a pQTL can reflect a variant that changes the binding site of the reagent rather than the amount of protein (an *epitope effect*). For single proteins at very low concentrations, *digital* immunoassays trap individual molecules in femtolitre wells and count them, which reaches sub-femtomolar sensitivity (Rissin et al. 2010).

8.6.3 Multiplexed imaging

Standard immunofluorescence measures a few proteins at a time. **Highly multiplexed imaging** measures dozens of proteins in the same tissue section, at single-cell resolution. In CODEX, antibodies are labelled with DNA barcodes that are read out in repeated cycles of fluorescent hybridisation (Goltsev et al. 2018). In **imaging mass cytometry**, antibodies are labelled with metal isotopes, and the tissue is ablated spot by spot with a laser and analysed by MS (Giesen et al. 2014). These methods are the protein counterparts of spatial transcriptomics (Section 7.6), and are analysed with similar computational methods: cell segmentation, cell-type assignment and analysis of spatial neighbourhoods.

8.6.4 Towards protein sequencing

All the methods above infer the identity of a protein indirectly. Several technologies are being developed to sequence single protein molecules directly, analogous to DNA sequencing. They include nanopores that read the current as a peptide passes through, fluorescent labelling of specific amino acids followed by sequential degradation, and binders that recognise the N-terminal amino acid (Alfaro et al. 2021). If successful, they would make it possible to count individual proteoforms with single-molecule sensitivity.

8.6.5 Orthogonal validation

Every protein measurement relies on a model of recognition: antibodies recognise epitopes, and MS recognises peptides through a search and inference pipeline. Validation is most informative when it tests the weakest link of the discovery method with an assay that does not share it. A second antibody against the same epitope is not independent of the first. A rough ordering of increasingly independent evidence is:

1. internal evidence from the MS data: co-eluting fragment ions, mass error, retention time and consistent peptides;
2. replication of the whole process (new extraction and digestion, or a new batch), not just a repeated injection;
3. targeted confirmation, for example PRM or SRM with heavy synthetic peptides, or comparison with the spectrum of a synthetic peptide;
4. biological tests: knockout or knockdown of the gene, perturbation of the modifying enzyme, or an expected localisation or time course;
5. replication in an independent cohort, laboratory or assay principle.

The appropriate level depends on the claim. A Western blot adds the size of the detected protein, which can show cleavage or a major isoform, and imaging adds localisation and cell-to-cell variation, but each needs its own controls. A biomarker intended for clinical use additionally needs calibration, defined detection limits, blinded samples and validation in independent cohorts.

See also: further explanation in Section D.8.5; related Wikipedia articles in Section E.8.5.

i Sources

Parts of this section are adapted from the Wikipedia articles Human Protein Atlas, Immunohistochemistry, Proximity extension assay and Aptamer (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Proteomics and Mass-Spectrometric Inference" (CC BY 4.0), used with permission.

8.7 Repositories and standards

Proteomics data are large and complex, and their reuse depends on public archives and common formats. As in genomics, depositing raw data has become a condition of publication.

- **PRIDE**, at EMBL-EBI, is the largest repository of MS proteomics data (Perez-Riverol et al. 2021, 2024).
- **ProteomeXchange** is a consortium of repositories, including PRIDE, PeptideAtlas, MassIVE, jPOST and iProX, which coordinates submission and gives each data set a stable identifier (Vizcaíno et al. 2014; Deutsch et al. 2022).
- **PeptideAtlas** reprocesses public data sets with a uniform pipeline and strict error control, and provides a compendium of observed peptides (Desiere 2006).

The Human Proteome Organization (HUPO) **Proteomics Standards Initiative** develops open data formats: **mzML** for raw spectra (Martens et al. 2011), mzIdentML for identifications and mzTab for summarised results. MIAPE (Minimum Information About a Proteomics Experiment) defines what should be reported, analogous to MIAME for microarrays (Taylor et al. 2007).

Public data are used for reanalysis, for building spectral libraries and for training the deep-learning models of Section 8.3. They also underlie the **Human Proteome Project**, launched by HUPO in 2010 to identify at least one protein product of every human protein-coding gene (Legrain et al. 2011). Two draft maps of the human proteome were published in 2014 (Kim et al. 2014; Wilhelm et al. 2014), but reanalysis showed that some of their claims, such as many proteins from non-coding regions, did not survive stringent error control (Ezkurdia, Vázquez, et al. 2014), an illustration of why protein-level FDR matters (Section 8.4). By 2023, credible MS or other protein evidence existed for about 93% of the predicted human proteins (Omenn et al. 2024). The remaining "missing proteins" are often expressed only in particular tissues or developmental stages, or are hard to detect by MS.

See also: further explanation in Section D.8.6; related Wikipedia articles in Section E.8.6.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Proteomics Identifications Database, PeptideAtlas, Proteomics Standards Initiative and Human Proteome Project (CC BY-SA 4.0, retrieved 4 October 2026).

8.8 Key concepts

A checklist for revision.

- **History:** 2D gels; soft ionisation (ESI and MALDI); peptide mass fingerprinting; the proteome
- **Instruments:** m/z; quadrupole, ion trap, TOF, FT-ICR and Orbitrap; hybrid instruments (Q-TOF, Q-Orbitrap); ion mobility (TIMS) and PASEF; resolution, mass accuracy, charge states and isotope envelopes; m/z vs. neutral mass and charge-state deconvolution
- **Strategies:** bottom-up (shotgun) vs. middle-down vs. top-down proteomics and proteoforms; loss of connectivity between modifications after digestion; LC-MS/MS; MS1 and MS2; CID/HCD and b/y ions; ETD and c/z ions
- **Acquisition:** data-dependent (top N, dynamic exclusion, missing values) vs. data-independent (SWATH) vs. targeted (SRM/PRM); duty cycle, DIA window width and cycle time
- **Quantification:** label-free (intensity, spectral counting, MaxLFQ); SILAC; isobaric tags (iTRAQ, TMT) and ratio compression; absolute quantification (AQUA, iBAQ); why none of these are molecule counts

- **Design and quality control:** SILAC label swaps; TMT plexes, pooled bridging channels, MS3 and confounding with plex; run order and drift; blanks, carryover, pooled QC and process standards; biological vs. technical replicates
- **Database search:** candidate peptides and precursor mass; theoretical spectra; SEQUEST XCorr, Mascot, Andromeda/MaxQuant, MS-GF+; search scores are not probabilities of correctness; the search space; open search (MSFragger); proteogenomics and separate FDR for novel peptides
- **Terminology:** precursor, peptide ion, peptide sequence, PSM, protein group
- **Spectral libraries** and library search
- **De novo sequencing:** spectrum graphs and dynamic programming; PEAKS; deep learning (DeepNovo, Casanovo, InstaNovo); uses (antibodies, immunopeptidomics, unsequenced organisms)
- **Spectrum and retention-time prediction:** Prosit, MS²PIP, AlphaPeptDeep; rescoring; predicted libraries; peptide-centric DIA analysis (OpenSWATH, DIA-NN)
- **Error control:** target–decoy approach; q-values and posterior error probabilities; PeptideProphet; Percolator and semi-supervised learning; PSM-, peptide- and protein-level FDR (always state the level); concatenated target–decoy competition; FDR as a property of a list, not of one PSM; subgroup FDR
- **Protein inference:** shared peptides; parsimony and protein groups; razor peptides; ProteinProphet; why peptides rarely identify a single protein; shared peptides and isoform-level quantification
- **Quantitative statistics:** summarisation of peptides; missing values (censored and random) and imputation; match between runs; MSstats and limma; a multiplicative model of peptide intensity; normalisation and its assumptions; identification FDR vs. differential-abundance FDR
- **PTMs:** phosphoproteomics, enrichment and site localisation (localisation probability vs. peptide FDR); identification, localisation, abundance and occupancy as separate questions; normalising sites to protein abundance; ubiquitination, acetylation and glycosylation
- **Interaction proteomics:** AP-MS and its contaminants (SAINT, CRAPome); BioPlex; proximity labelling (BioID, APEX); co-fractionation MS; hu.MAP 2.0 and 3.0 and how they are built
- **Structural MS:** cross-linking MS and its search problem; HDX-MS; native MS; LiP-MS and thermal proteome profiling
- **Spatial proteomics:** organelle profiling; the HPA subcellular atlas; MS imaging; deep visual proteomics; single-cell proteomics
- **Non-MS proteomics:** antibodies and their validation; antibody detection vs. MS identification; the Human Protein Atlas; proximity extension assays (Olink) and aptamers (SomaScan); pQTLs and epitope effects; multiplexed imaging (CODEX, imaging mass cytometry); digital immunoassays; single-molecule protein sequencing; orthogonal validation and its levels
- **Repositories and standards:** PRIDE, ProteomeXchange, PeptideAtlas; mzML, mzIdentML, mzTab, MIAPE; the Human Proteome Project and missing proteins

8.9 Recommended reading

These papers are recommended for reading in full:

- Aebersold and Mann (2016) — mass-spectrometric exploration of proteome structure and function; the field in one review.
- Cox and Mann (2008) — MaxQuant; how a complete shotgun proteomics pipeline identifies and quantifies proteins.
- Elias and Gygi (2007) — the target–decoy strategy; how false discoveries are controlled in proteomics.
- Ludwig et al. (2018) — data-independent acquisition (SWATH-MS), a tutorial.
- Yilmaz et al. (2024) — Casanovo; de novo peptide sequencing as translation with a transformer.

- Uhlén et al. (2015) — the tissue-based map of the human proteome (Human Protein Atlas).

Titles, abstracts, links and the full texts are in Section A.8; a longer list of further papers is in Section B.8.

8.10 Background reading (open access)

- **Proteomics (Wikipedia)** — Proteomics is the large-scale study of proteins. The proteome is the entire set of proteins produced or modified by an organism or system. Proteomics is an interdisciplinary field that covers the exploration of proteomes from the overall level of protein composition, structure, and activity.
  https://en.wikipedia.org/wiki/Proteomics
- **Shotgun proteomics (Wikipedia)** — Shotgun proteomics refers to the use of bottom-up proteomics techniques in identifying proteins in complex mixtures using a combination of high performance liquid chromatography combined with mass spectrometry.
  https://en.wikipedia.org/wiki/Shotgun_proteomics
- **Mass spectrometry (Wikipedia)** — Mass spectrometry (MS) is an analytical technique that is used to measure the mass-to-charge ratio of ions. The results are presented as a mass spectrum, a plot of intensity as a function of the mass-to-charge ratio.
  https://en.wikipedia.org/wiki/Mass_spectrometry
- **Tandem mass spectrometry (Wikipedia)** — Tandem mass spectrometry, also known as MS/MS or MS2, is a technique in instrumental analysis where two or more stages of analysis using one or more mass analyzers are performed with an additional reaction step in between these analyses to increase their abilities to analyse chemical samples.
  https://en.wikipedia.org/wiki/Tandem_mass_spectrometry
- **Peptide mass fingerprinting (Wikipedia)** — Peptide mass fingerprinting (PMF), also known as protein fingerprinting, is an analytical technique for protein identification in which the unknown protein of interest is first cleaved into smaller peptides, whose absolute masses can be accurately measured with a mass spectrometer such as MALDI-TOF or ESI-TOF.
  https://en.wikipedia.org/wiki/Peptide_mass_fingerprinting
- **De novo peptide sequencing (Wikipedia)** — In mass spectrometry, de novo peptide sequencing is the method in which a peptide amino acid sequence is determined from tandem mass spectrometry.
  https://en.wikipedia.org/wiki/De_novo_peptide_sequencing
- **Data-independent acquisition (Wikipedia)** — In mass spectrometry, data-independent acquisition (DIA) is a method of molecular structure determination in which all ions within a selected m/z range are fragmented and analyzed in a second stage of tandem mass spectrometry.
  https://en.wikipedia.org/wiki/Data-independent_acquisition
- **SILAC (Wikipedia)** — Stable isotope labeling by amino acids in cell culture (SILAC) is a technique based on mass spectrometry that detects differences in protein abundance among samples using non-radioactive isotopic labeling. It is a popular method for quantitative proteomics.
  https://en.wikipedia.org/wiki/Stable_isotope_labeling_by_amino_acids_in_cell_culture
- **Isobaric labeling (Wikipedia)** — Isobaric labeling is a mass spectrometry strategy used in quantitative proteomics. Peptides or proteins are labeled with chemical groups that have nominally identical mass (isobaric), but vary in terms of distribution of heavy isotopes in their structure.
  https://en.wikipedia.org/wiki/Isobaric_labeling
- **Human Protein Atlas (Wikipedia)** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.
  https://en.wikipedia.org/wiki/Human_Protein_Atlas

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

8.11 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

8.12 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapter "Proteomics and Mass-Spectrometric Inference" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

8.12.1 Design a proteomics experiment

You want to discover differences in protein abundance between patients who respond to a treatment and patients who do not. How would you design the study?

Extensions, solution outline and common errors

Extensions:

1. What is the independent biological replicate?
2. Would you use DDA, DIA, TMT or a targeted panel?
3. How would you handle preparation batches and run order?
4. What controls would you add, and at which stage?
5. What can you conclude if a protein is not identified in the non-responders?

Solution outline:

- Define what is to be compared, the population, the clinical covariates, the expected effect size, and whether the study is for discovery or confirmation.
- Treat patients, not LC fractions, peptides, spectra or repeated injections, as the independent units, and plan the number of patients from realistic biological variability.
- Balance responders and non-responders across sample collection, preparation batches, acquisition batches and TMT plexes. Randomise the run order and keep paired or matched structure.
- Label-free DIA suits a large cohort that needs reproducible, broad coverage; TMT is an option when multiplexing improves throughput and the plexes can be balanced; targeted assays are for a predefined panel.
- Add process standards before digestion, system standards before and during acquisition, pooled QC samples throughout the run and blanks to detect carryover.
- Decide the rules for exclusion, normalisation, missing values, the statistical model and multiple testing in advance, and keep an independent cohort for validation.
- Non-detection is ambiguous: the protein may be absent, or lost in sampling, extraction or acquisition, below the dynamic range, or not identified.

Common errors:

- Calling repeated injections, or TMT channels from one sample, independent replicates.
- Confounding response with batch and proposing correction afterwards as a cure.
- Choosing the platform only from the number of proteins it claims to identify.
- Treating a missing value as zero expression.
- Validating only in the discovery cohort.

Connections: Section 8.2 (experimental design and quality control); batch effects and design matrices in Section 7.4; validation and multiple testing in 4.

8.12.2 LC-MS/MS and acquisition strategies

Explain how a peptide becomes an identification in LC-MS/MS, then compare DDA and DIA.

Extensions, solution outline and common errors

Extensions:

1. Why is chromatography necessary?
2. What information is in MS1 and MS2?
3. Why does DDA give missing values that vary from run to run?
4. Why are DIA spectra harder to analyse?
5. Where do targeted PRM and SRM fit?

Solution outline:

- Reversed-phase LC reduces the complexity of what enters the instrument at any moment and spreads peptides over the retention time; electrospray creates gas-phase ions with one or more charges.
- MS1 measures the precursor m/z, the isotope envelope (and hence the charge) and the intensity. MS2, after fragmentation, gives b- and y-ion series that carry sequence information.
- DDA selects a limited number of precursors, usually the most intense. Dynamic exclusion improves coverage, but the competition between precursors differs between runs, so different peptides are selected.
- DIA fragments all precursors in a series of windows, which is more reproducible but mixes fragments from several peptides. The analysis uses co-elution, mass accuracy, measured or predicted spectra, retention times and scores that account for interference.
- Narrower windows reduce interference, but each chromatographic peak must still be sampled several times for quantification, which limits the number of windows.
- PRM and SRM monitor predefined precursors and fragments and are used for verification, especially with heavy internal standards.

Common errors:

- Saying that the instrument reads the peptide sequence directly.
- Equating high precursor mass accuracy with a unique identification.
- Calling DIA “unbiased”: ionisation, mass range, windows and libraries still select what is seen.
- Claiming that DIA has no missing values.
- Ignoring chromatography and the duty cycle.

Connections: Section 8.2; peptide-centric DIA analysis in Section 8.3; AP-MS in Section 8.5 and Section 11.2.

8.12.3 Database search, de novo sequencing and FDR

How does a database search assign a tandem mass spectrum to a peptide, and how would you know that the accepted matches are reliable?

Extensions, solution outline and common errors

Extensions:

1. What changes when you allow many variable modifications?
2. How does target–decoy competition estimate the FDR?
3. What does a 1% PSM-level FDR mean?
4. When would you use de novo sequencing?
5. How would you validate a novel peptide?

Solution outline:

- Candidate peptides are generated from a specified protein database with rules for the enzyme, missed cleavages, charge, mass tolerance and modifications; their theoretical or predicted fragments are compared with the observed spectrum.
- A larger search space contains more competing candidates, increases the chance of high-scoring random matches and can reduce the number of identifications at a fixed FDR. A peptide that is not in the database can never be found by a standard search.
- In concatenated target–decoy competition, wrong matches are assumed to hit targets and decoys equally often, so the decoys accepted above a threshold estimate the false targets.
- A 1% PSM FDR is the expected fraction of wrong PSMs in the accepted list. It is not a 1% error probability for each PSM, and it does not imply a 1% FDR for peptides, sites or proteins.
- De novo sequencing is used when no suitable database exists, for antibodies and immunopeptides, or to find variant and novel peptides. Its large search space, incomplete fragmentation and isobaric residues require careful assessment of confidence.
- A novel peptide can be validated with the spectrum of a synthetic peptide, targeted MS, genomic or transcript evidence, independent samples, and a separate FDR for novel peptides.

Common errors:

- Defining the FDR as the probability that a particular result is false.
- Estimating the PSM-level FDR and reporting it as the protein-level FDR.
- Adding every possible PTM “to be safe”.
- Treating a high score from a deep-learning de novo model as proof of a new gene.
- Forgetting contaminants and the database version.

Connections: Section 8.3; Section 8.4; database size and statistical significance in sequence searching (Section 1.4); proteogenomics and gene annotation (Section 5.6).

8.12.4 Compare quantitative strategies

Compare label-free quantification, SILAC and TMT for a study of a perturbation with several conditions.

Extensions, solution outline and common errors

Extensions:

1. Why does early mixing help SILAC?
2. Why can TMT ratios be compressed?
3. How would you connect several TMT plexes?
4. What does normalisation assume?
5. Can the intensities of two different peptides be compared directly?

Solution outline:

- Label-free quantification compares peak areas between separate runs and scales to many samples, but retention-time alignment, drift, batches and missing values must be handled.
- SILAC labels proteins metabolically; light and heavy peptides co-elute, and mixing early means that later handling affects both equally. It suits cell culture, and needs checks of label incorporation and arginine-to-proline conversion, and often label swaps.
- TMT labels peptides with isobaric tags and quantifies reporter ions after fragmentation. It multiplexes samples and reduces missing values within a plex.
- Co-isolated peptides contribute reporter ions and pull ratios towards one. Narrow isolation, better separation, filtering of spectra with interference, or MS3 reduce this compression.
- Balance the conditions across plexes and include a pooled reference channel in each. Compare mainly within plexes and include the plex in the statistical model; a reference channel cannot rescue a condition that is confounded with a plex.
- Normalisation assumes that most proteins are unchanged, or relies on an internal standard or another reference. A global biological shift violates median scaling.
- Peptides differ in ionisation and recovery, so only changes in the same peptide between samples can be compared reliably.

Common errors:

- Saying that TMT gives absolute quantities.
- Treating the pooled reference channel as an independent biological sample.
- Claiming that SILAC removes variation that arose before labelling or mixing.
- Normalising until all groups overlap.
- Combining modified, shared and discordant peptides without inspection.

Connections: Section 8.2; normalisation and peptide intensities in Section 8.4; normalisation and linear models for RNA-seq in Section 7.4.

8.12.5 Protein inference and proteoforms

You identify three peptides from a gene: two are shared by two splice isoforms, and one is unique to the longer isoform. What can you report?

Extensions, solution outline and common errors

Extensions:

1. What if only the shared peptides were observed?
2. What does parsimony do?

3. How does the protein-level FDR differ from the PSM-level FDR?
4. Can the shared peptides be used to quantify the longer isoform?
5. When would top-down proteomics help?

Solution outline:

- Check that the "unique" peptide is unique in the exact database that was searched, and that it is accepted at an appropriate error rate.
- With reliable unique evidence, the longer isoform is supported, given that database. The shared peptides support the group of both isoforms.
- With only shared peptides, report a protein group; do not choose one isoform because it is the canonical one or expected to be expressed.
- Parsimony finds the smallest set of proteins that explains the observed peptides. It is a reporting rule, not evidence that the smallest set is what is present.
- The protein-level FDR must be estimated after PSMs and peptides have been combined into proteins, for example with the picked target–decoy approach.
- The abundance of shared peptides cannot be assigned to one isoform without further assumptions. Quantify the group, or use isoform-specific peptides and check the sensitivity of the conclusions.
- Top-down proteomics keeps the intact combination of sequence and modifications, but with lower depth and greater technical difficulty.

Common errors:

- Calling a gene-level group a specific isoform.
- Counting several spectra of one peptide as independent observations of the sequence.
- Treating parsimony as a law of nature.
- Assuming that a peptide unique in one database is unique in all proteomes and variants.
- Using RNA levels to override ambiguous peptide evidence.

Connections: Section 8.4; bottom-up and top-down proteomics in Section 8.2; alternative splicing in Section 3.5; gene models and annotation in Section 5.6.

8.12.6 PTMs and site occupancy

A phosphopeptide increases fourfold after stimulation. What does this result mean, and what further evidence do you need?

Extensions, solution outline and common errors

Extensions:

1. How do you distinguish identification from site localisation?
2. Could a change in the total amount of protein explain the result?
3. How would you estimate occupancy?
4. Why is the confidence in the site separate from the peptide FDR?
5. How would you test whether the site is a direct substrate of a kinase?

Solution outline:

- The direct result is that the signal of an enriched, modified peptide changed, given its identification, localisation, normalisation and acquisition.

- A peptide can be identified with confidence while the phosphate is ambiguous between neighbouring residues; site-determining fragment ions and a localisation score (such as the Ascore) are needed.
- Compare with a measurement of the unenriched proteome. A proportional change in the protein can explain the change in the phosphopeptide; an unchanged protein level makes a change in modification more plausible.
- Occupancy is the number of modified molecules divided by the number of modified plus unmodified molecules. It requires both forms and a correction for their different responses; a fold change in enriched material is not occupancy.
- Localisation errors are controlled separately from identification errors, and ambiguous sites should be reported as such.
- Direct substrates can be tested with kinase inhibition or knockout, time courses, in vitro kinase assays, sequence motifs, targeted MS with a synthetic phosphopeptide and mutation of the site. None of these is sufficient on its own.

Common errors:

- Calling every regulated site a direct substrate of the kinase.
- Confusing a localisation probability with the PSM-level FDR.
- Ignoring variation in enrichment and the total amount of protein.
- Equating phosphopeptide intensity with occupancy.
- Using a site-specific antibody without controls for specificity.

Connections: Section 8.5; protein function in Chapter 10; signalling networks in Section 11.3.

8.12.7 Missing values, QC and differential abundance

Half the values for a low-abundance protein are missing in the controls but observed in the treated samples. How would you analyse and interpret this?

Extensions, solution outline and common errors

Extensions:

1. Why not replace the missing values with the minimum intensity?
2. How can the acquisition strategy affect missing values?
3. Which quality-control plots would you inspect?
4. What is the number of replicates in the statistical test?
5. How would you validate apparent on/off expression?

Solution outline:

- Look at the peptide-level observations by sample, intensity, condition, batch and run order, and at the raw chromatograms and identification scores, before working with an imputed protein table.
- Consider detection that depends on abundance, competition between precursors in DDA, failed peak integration, batches, and real presence in only one condition. Several mechanisms may act together.
- Replacing every missing value with a fixed low value can create an artificial difference between groups and hides the uncertainty. Prefer a model that treats low values as censored, sensitivity analyses, or a presence/absence analysis with explicit assumptions.

- Inspect the coefficient of variation of pooled QC samples, retention-time drift, spray stability, mass errors, blanks, digestion metrics, normalisation factors, agreement between peptides and sample identity.
- Independent donors or cultures define the replicates; peptides and spectra are nested measurements within them.
- An on/off claim can be confirmed with targeted MS and heavy peptides, dilution series to establish the detection limit, process blanks and independent biological samples.

Common errors:

- Treating all missing values as censored, or all as random.
- Imputing before splitting the data into training and test sets for a classifier.
- Counting peptide observations as independent replicates.
- Removing an outlier because it weakens the significance.
- Equating "below the detection limit" with zero molecules.

Connections: Section 8.4; quality control in Section 8.2; data leakage and validation in 4.

8.12.8 Choose an orthogonal validation method

When should a discovery by MS be validated with targeted MS, a Western blot, ELISA, flow cytometry, a proximity extension assay or an aptamer panel?

Extensions, solution outline and common errors

Extensions:

1. What does "orthogonal" mean here?
2. How would you validate a claim about a specific isoform?
3. Which controls does an antibody-based assay require?
4. Why might MS and an affinity assay disagree?
5. Which methods can give absolute concentrations?

Solution outline:

- Choose a method that tests the weakest link of the claim. PRM with heavy peptides tests the identity of the peptide and its quantity; a Western blot adds the apparent size; ELISA gives sensitive targeted concentrations; flow cytometry and imaging add cell type and location; PEA and aptamer panels add multiplexed affinity measurements.
- Orthogonal means using recognition and error mechanisms that are at least partly independent, not just analysing the same spectra with different software.
- An isoform-specific claim can be validated with an isoform-unique peptide, intact-protein (top-down) evidence, isoform-specific genetic perturbation, or an antibody whose epitope distinguishes the isoforms.
- Antibodies need tests such as knockout or knockdown, the expected size, linear response on dilution, absence of saturation, and reproducibility between lots (Section 8.6).
- Disagreement can arise because the assays recognise different epitopes or proteoforms, because variants change binding (epitope effects), because complexes mask epitopes, or because one platform suffers from interference.
- Absolute concentration requires calibration and recovery controls, as in targeted MS with isotope-labelled standards or a calibrated immunoassay. Label-free intensities, TMT ratios, PEA units and aptamer signals are usually relative.

Common errors:

- Assuming that every antibody assay is independent and specific.
- Taking disagreement between platforms as proof that one result is false.
- Confusing a precise relative signal with an absolute concentration.
- Validating only on the same samples.
- Using RNA levels as direct validation of protein levels.

Connections: Section 8.6; mRNA and protein levels in Section 7.5; interaction evidence in Section 11.2.

9 Protein structure

> **From the exam description**
>
> Knowledge of protein structure including domains and different methods to predict protein structure.

9.1 Overview

Protein structure is a central topic of the exam and of bioinformatics. A protein's function depends on its three-dimensional structure, and predicting structure from sequence was a grand challenge of computational biology for fifty years. In 2020, deep learning made it possible to predict the static structure of many single-chain proteins, in particular globular domains, with close to experimental accuracy. Conformational ensembles, disordered regions, the effects of mutations, many complexes, and states that depend on ligands, ions or conditions remain open problems. The field is still moving very fast, from single chains to complexes, ligands, dynamics and design, and this chapter tries to be as up to date as possible. It covers:

- CASP, the blind experiment that has measured progress since 1994 (Section 9.2);
- background: amino acids, thermodynamics and binding, and enzymes (Section 9.3);
- the levels of protein structure and the experimental methods: X-ray crystallography, NMR, cryo-EM and integrative modelling (Section 9.4);
- protein folding: Anfinsen's dogma, stability, Levinthal's paradox and funnels, and misfolding, amyloid and prions (Section 9.5);
- structure comparison, what it has revealed about structure evolution, and the structure databases (Section 9.6 and Section 9.7);
- the history of structure prediction, from secondary structure (and how it is assigned from coordinates) and homology modelling with loops and side chains to AlphaFold3 and co-folding models, its limits, including intrinsically disordered proteins, and how to check a predicted model (Section 9.8);
- protein language models and structure (Section 9.9);
- membrane proteins (Section 9.10);
- molecular dynamics, Monte Carlo sampling, convergence and computational cost, optimisation and energy functions (force fields and statistical potentials), normal modes and docking (Section 9.11);
- protein design (Section 9.12).

The prediction of complexes is continued in Chapter 11, and the use of structure to predict function in Chapter 10.

9.2 CASP: a blind test of structure prediction

Much of the history of structure prediction can be told through one experiment. The **Critical Assessment of protein Structure Prediction** (CASP) was started in 1994 by John Moult and colleagues

(Moult et al. 1995). Before CASP, methods were usually tested on structures that were already known, often by the developers themselves, and claims of success were hard to evaluate. CASP introduced a simple principle: predictions must be made **blind**. In the original design, the organisers collected, every second year, sequences of proteins whose structures were about to be determined experimentally but were not yet public. Today the targets come from several routes, mostly unpublished structures contributed by experimental groups, and include not only single proteins but also assemblies, RNA, protein–ligand complexes and other categories. The principle is unchanged: predictors submit models before the structures are released, and independent assessors compare the models with the experimental structures. CASP has run every two years since then, and is now organised by the Protein Structure Prediction Center at UC Davis.

9.2.1 Categories and metrics

Targets were originally classified by how much information was available: *comparative modelling* targets, with a clear template in the PDB; *fold recognition* targets, with a template that is hard to find; and *free modelling* (originally *ab initio*) targets, with no known related structure. Over the years CASP added categories for residue contacts, model quality estimation, refinement, complexes (assembly), ligands, RNA, and conformational ensembles. Two community experiments complement it: **CAPRI** (Critical Assessment of PRedicted Interactions) for protein docking, running since 2001 (Janin et al. 2003), and **CAMEO**, which evaluates servers continuously every week on newly released PDB entries (Haas et al. 2017).

The experimental structure used as reference is itself a model of one state; for many proteins the native state is an ensemble of conformations (Section 9.5), and flexible regions may be absent or poorly defined. Comparing a model with the experimental structure requires measures that are robust to errors in part of the model:

- **RMSD** after optimal superposition is the classical measure, but it is dominated by the worst-predicted regions, so a model with a correct core and one misplaced loop gets a poor score (Section 9.6).
- **GDT_TS** (global distance test) is the average percentage of residues that can be superimposed within 1, 2, 4 and 8 Å of the experimental positions, maximised over many superpositions (Zemla 2003). It ranges from 0 to 100, and values above about 90 are considered competitive with experimental accuracy.
- The **TM-score** weights distances so that it is independent of protein length, and values above 0.5 indicate the same fold (Zhang and Skolnick 2004).
- **lDDT** (local distance difference test) does not superimpose at all. It measures the fraction of interatomic distances in the reference that are preserved in the model within given tolerances (Mariani et al. 2013). Because it is local, it is robust to domain movements, and its predicted version, pLDDT, is the confidence measure of AlphaFold (Section 9.8).
- For complexes, **DockQ** combines the fraction of native contacts and interface and ligand RMSDs into a single score between 0 and 1 (Basu and Wallner 2016).

9.2.2 What CASP has shown

The CASP record documents both slow progress and sudden jumps (Kryshtafovych et al. 2021):

- **1994–2004.** Comparative modelling was useful when templates were close, and fold recognition by threading and profile–profile alignment improved steadily. Ab initio predictions were mostly wrong, until fragment-assembly methods, especially Rosetta, produced the first correct folds of small proteins in CASP3–CASP5.

- **2006–2016.** Progress was incremental. Consensus methods (meta-servers) and model quality estimation became important. From CASP11 (2014) the *coevolution* methods made correct predictions of some large proteins without templates possible, when deep alignments were available.
- **2018 (CASP13).** Deep-learning prediction of residue distances, by several groups and most strikingly by DeepMind's first AlphaFold (Senior et al. 2020), gave a large jump for free-modelling targets.
- **2020 (CASP14). AlphaFold2** produced models with a median GDT_TS of about 92 across all targets, comparable to experimental accuracy for most single domains (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021; Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Silver, et al. 2021). The organisers declared that the prediction of single-domain protein structures was largely solved; this referred to static structures of single chains, not to ensembles, complexes or ligands.
- **2022 (CASP15)** and **2024 (CASP16).** Almost all groups built on AlphaFold2 or, in CASP16, on the publicly available AlphaFold3 server, combined with better alignments, extensive sampling and model selection. AlphaFold3 was not a new entrant in the way AlphaFold2 was in CASP14: it was released during CASP16, and many participants used its public server as a component of their pipelines. In CASP16 no single-domain target was predicted with the wrong fold. The remaining challenges are large complexes, flexible and interaction-dependent regions, RNA, ligands and alternative conformations (Kryshtafovych et al. 2023; Yuan et al. 2025).

CASP's success inspired similar blind assessments in other fields, such as CAPRI for docking, CAFA for function prediction (Chapter 10) and CAMI for metagenomics (Section 5.7). Its lesson for the exam is general: a method is only as good as its performance on data it has not seen, a theme that returns in Chapter 12. Outside blind tests, benchmarks of structure predictors must control for overlap between the test proteins and the templates or training structures the method has access to; such leakage inflates apparent accuracy (4).

See also: further explanation in Section D.9.1; related Wikipedia articles in Section E.9.1.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles CASP, Global distance test and Template modeling score (CC BY-SA 4.0, retrieved 4 October 2026).

9.3 Background: amino acids, energetics and enzymes

Examiners expect you to know the chemistry behind the methods in this chapter, at the level of an undergraduate biochemistry course. This section summarises it. The corresponding background on nucleic acids, genes and cells is given in Section 6.2.

9.3.1 Amino acids

Proteins are linear polymers of 20 standard amino acids, joined by **peptide bonds**. The peptide bond is planar, so the conformation of the backbone is described by two torsion angles per residue, ϕ and ψ (Section 9.4). The amino acids differ in their side chains, and their properties explain much of protein structure and evolution:

- **hydrophobic** side chains (Ala, Val, Leu, Ile, Met, Phe, Trp) are buried in the core of soluble proteins and face the lipids in membrane proteins;
- **polar** side chains (Ser, Thr, Asn, Gln, Tyr) form hydrogen bonds and are common on the surface;
- **charged** side chains, positive (Lys, Arg, His at low pH) and negative (Asp, Glu), form salt bridges and are found on the surface and in active sites;
- special cases: **glycine**, without a side chain, is flexible and common in turns; **proline**, whose side chain is bonded to the backbone, is rigid and breaks helices; **cysteine** can form disulfide bonds.

Substitution matrices such as PAM and BLOSUM (Section 1.3) reflect these properties: substitutions between similar amino acids, such as Ile and Val, are frequent in evolution, while substitutions involving Cys, Trp or Gly are rare.

9.3.2 Thermodynamics and binding

Whether a protein folds, and how strongly molecules bind, are questions of **free energy**, $\Delta G = \Delta H - T\Delta S$, a balance of enthalpy (interactions) and entropy (disorder). The main forces are the **hydrophobic effect**, the main driver of folding, which is largely entropic: non-polar groups order the surrounding water, and burying them releases it; **hydrogen bonds**; **electrostatic** interactions; and **van der Waals** contacts. Folded proteins are only marginally stable (Section 9.5).

Binding of a ligand L to a protein P is described by the **equilibrium dissociation constant**, $K_d = [\mathrm{P}][\mathrm{L}]/[\mathrm{PL}]$, which has units of concentration and equals the free ligand concentration at which half of the protein is bound. K_d is an equilibrium constant, not a rate; it equals the ratio of the dissociation and association rate constants, $k_{\mathrm{off}}/k_{\mathrm{on}}$. It is not itself a free energy, but it is related to the **standard binding free energy** by

$$\Delta G^{\circ}_{\mathrm{bind}} = RT \ln \frac{K_d}{c^{\circ}},$$

where $c^{\circ} = 1$ M is the standard-state concentration (equivalently, $\Delta G^{\circ}_{\mathrm{bind}} = -RT \ln K_a$ with the dimensionless association constant $K_a = c^{\circ}/K_d$). The value of ΔG° therefore depends on the chosen standard state. A K_d of 1 nM corresponds to about -12.3 kcal/mol at 25 °C, and a tenfold lower K_d (tighter binding) to about 1.4 kcal/mol more favourable binding. Typical K_d values range from millimolar for weak, transient interactions to picomolar for antibodies and tight complexes. Predicting binding affinities is one of the hardest problems in computational biology (Section 9.11).

9.3.3 Enzymes

Enzymes accelerate reactions by lowering the free energy of the transition state, without changing the equilibrium. Many reaction rates are increased by factors of 10^{10} or more. For a simple enzyme, the rate depends on the substrate concentration [S] by the **Michaelis–Menten** equation,

$$v = \frac{V_{\max}[S]}{K_m + [S]}, \qquad V_{\max} = k_{\mathrm{cat}}[E]_{\mathrm{total}},$$

where K_m is the substrate concentration at half-maximal rate, and k_{cat} the turnover number, the number of reactions per enzyme per second. K_m is **not** in general the dissociation constant of the enzyme–substrate complex. For the scheme $\mathrm{E} + \mathrm{S} \rightleftharpoons \mathrm{ES} \rightarrow \mathrm{E} + \mathrm{P}$, with rate constants k_1, k_{-1} and k_{cat}, $K_m = (k_{-1} + k_{\mathrm{cat}})/k_1$, which equals $K_d = k_{-1}/k_1$ only when $k_{\mathrm{cat}} \ll k_{-1}$ (the rapid-equilibrium limit). A low K_m therefore does not by itself show tight substrate binding. k_{cat}/K_m measures catalytic efficiency. Inhibitors reduce activity, competitively by binding the active site or non-competitively elsewhere. Activity is **regulated**

by allosteric effectors, which bind away from the active site and shift the enzyme between conformations, and by post-translational modifications such as phosphorylation (Section 8.5). Enzymes are classified by the reactions they catalyse (EC numbers, Section 10.2), and their parameters are collected in BRENDA. Membranes and membrane proteins are described in Section 9.10, and metabolism in Section 11.3.

See also: further explanation in Section D.9.2; related Wikipedia articles in Section E.9.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Amino acid, Peptide bond, Dissociation constant, Enzyme and Michaelis–Menten kinetics (CC BY-SA 4.0, retrieved 4 October 2026). For more detail, see Alberts et al. (2002) or Clark et al. (2018).

9.4 Protein structure and how it is determined

A protein's function depends on its three-dimensional structure. This section introduces the vocabulary of protein structure and the experimental methods that produced the roughly 260,000 structures in the Protein Data Bank (in 2026) (Section 2.5).

9.4.1 Levels of protein structure

Protein structure is described at four levels:

- **Primary structure**: the amino acid sequence.
- **Secondary structure**: local, regular conformations stabilised by backbone hydrogen bonds. Linus Pauling predicted the **α-helix** and the **β-sheet** in 1951 from the geometry of the peptide bond, before any protein structure was known (Pauling et al. 1951). The allowed combinations of the backbone torsion angles ϕ and ψ are summarised in the **Ramachandran plot** (Ramachandran et al. 1963), still the first check of any structure model. Regions without regular structure are called loops or coils.
- **Tertiary structure**: the three-dimensional fold of a whole chain. Hydrophobic side chains are buried in a core, polar ones exposed to the solvent. Secondary structure elements pack against each other in a limited number of ways: helices pack at preferred angles determined by how their side-chain "ridges" fit into "grooves", and β-sheets are twisted and pack face to face or roll into barrels (Chothia et al. 1977). Common combinations of a few elements, such as the β-hairpin, the helix–turn–helix or the β-α-β unit, are called **supersecondary structures** or motifs.
- **Quaternary structure**: the assembly of several chains into a complex.

Larger proteins consist of **domains**, compact units of typically 50–250 residues that often fold independently and are reused in different combinations during evolution (Section 1.9 and Section 3.5). Domains are classified by their **fold**, the arrangement and connectivity of their secondary structure elements, and grouped into **superfamilies** when there is evidence of common ancestry. From the first structures, Levitt and Chothia defined four **structural classes**: all-α, all-β, α/β (alternating helices and strands, mostly parallel sheets) and α+β (segregated helices and mostly antiparallel sheets) (Levitt and Chothia 1976). Jane Richardson's drawings and taxonomy of structures, which introduced the ribbon diagram, shaped how structures are still visualised (Richardson 1981).

9.4.2 X-ray crystallography

The first protein structures, of myoglobin (Kendrew et al. 1958) and haemoglobin (PERUTZ et al. 1960), were solved by **X-ray crystallography** in Cambridge; Kendrew and Perutz received the Nobel Prize in Chemistry in 1962. In crystallography, a protein is crystallised, and the crystal diffracts an X-ray beam into a pattern of spots. The intensities of the spots give the amplitudes of the Fourier transform of the electron density, but not their phases (*the phase problem*). Phases are obtained from heavy-atom derivatives, from anomalous scattering, or, most often today, by **molecular replacement** with a similar known structure, increasingly a predicted model. An atomic model is then built into the electron density and refined.

The quality of a crystal structure is described by:

- the **resolution**, the smallest spacing of the diffraction data used. At 3 Å, the backbone and most side chains can be traced; at better than 1.5 Å, individual atoms are resolved;
- the **R-factor**, the disagreement between the observed amplitudes and those calculated from the model. Because a model can be over-fitted to the data, Axel Brünger introduced the **free R-factor**, computed on a small set of reflections (typically 5%) that is excluded from refinement (Brünger 1992). R_{free} is a cross-validation statistic, the same idea as a held-out test set in machine learning (Chapter 12). A large gap between R and R_{free} indicates over-fitting.

Crystallography gives the most accurate structures and remains the main method for drug design, but proteins must crystallise, and crystal contacts can distort flexible regions. More than 80% of PDB entries (as of 2026) were determined by X-ray crystallography.

9.4.3 NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy determines structures of proteins in solution. Kurt Wüthrich's group solved the first protein structure by NMR in 1985 (Williamson et al. 1985), and Wüthrich shared the 2002 Nobel Prize in Chemistry. NMR measures interactions between atomic nuclei, such as the nuclear Overhauser effect, which gives distances between protons closer than about 5 Å. A structure is calculated as an **ensemble** of models consistent with these restraints. NMR provides information about dynamics that crystallography cannot, but it is limited to relatively small proteins, typically below 30–50 kDa.

9.4.4 Cryo-electron microscopy

In **cryo-electron microscopy** (cryo-EM), purified molecules are frozen in a thin layer of vitreous ice and imaged with an electron microscope. Each image is very noisy, so hundreds of thousands of particle images are aligned, classified and averaged to reconstruct a three-dimensional density map. Richard Henderson used electron crystallography to determine the structure of bacteriorhodopsin, a membrane protein, in 1990 (Henderson et al. 1990). For decades single-particle cryo-EM gave only low resolution, until direct electron detectors and new image-processing software, with Bayesian methods for classification and reconstruction (Scheres 2012), produced the **resolution revolution** around 2013 (Kühlbrandt 2014). Jacques Dubochet, Joachim Frank and Richard Henderson received the 2017 Nobel Prize in Chemistry. Cryo-EM needs no crystals, works best for large complexes and membrane proteins, and images individual molecules rather than an ordered lattice. Classification of the particle images can therefore separate several conformational or compositional states from the same sample, and newer methods model continuous variability between them. The relative numbers of particles give a rough indication of how populated the states are, which connects experimental structure to dynamics (Section 9.11), although sample preparation and freezing can shift these proportions. Cryo-EM now

produces a large share of new structures, and **cryo-electron tomography** images complexes inside cells. In cryo-EM, resolution is estimated from the agreement between maps reconstructed from two independent halves of the data (the Fourier shell correlation), another form of cross-validation.

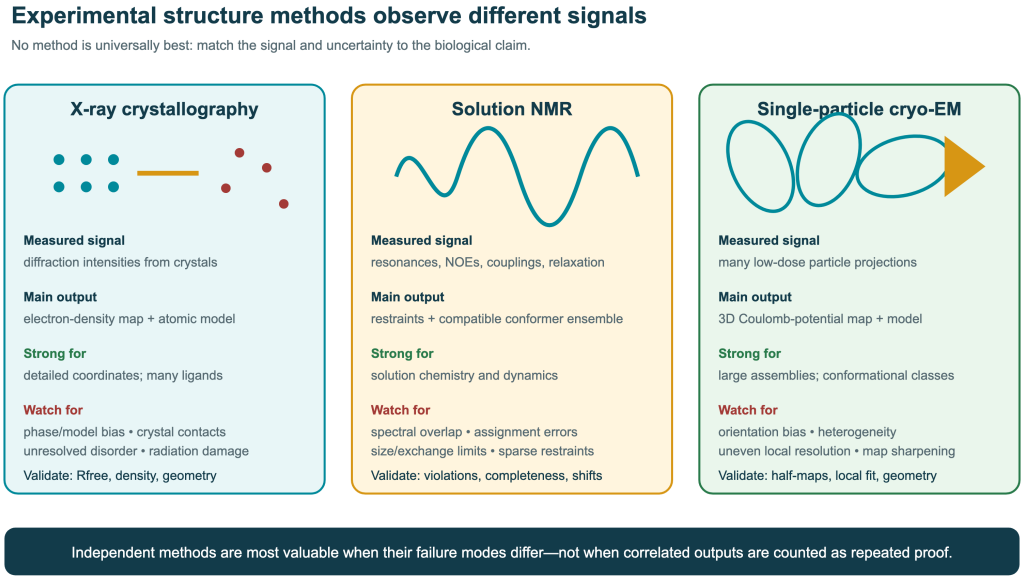


Figure 9.1: X-ray crystallography, solution NMR and single-particle cryo-EM compared: the signal each method measures, its main output, what it is strong for, its typical problems, and how its models are validated. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

9.4.5 Integrative and lower-resolution methods

Many systems are too large, flexible or heterogeneous for any one of the three main methods. **Integrative structural biology** combines several kinds of data, each too sparse to determine a structure alone, into models that satisfy all of them (Rout and Sali 2019). Common sources of restraints are:

- **small-angle X-ray scattering** (SAXS), which gives the overall size and shape of molecules in solution, averaged over all orientations and conformations;
- **cross-linking mass spectrometry**, which shows which residues are within the reach of a chemical cross-linker (Section 11.2);
- **hydrogen–deuterium exchange MS**, which shows which parts of the backbone are protected from the solvent, for example by an interface (Section 8.5);
- FRET and electron paramagnetic resonance, which measure distances between labelled sites, together with mutagenesis, low-resolution cryo-EM or tomography maps, and atomic structures or predicted models of the components.

A landmark example is the yeast nuclear pore complex, a ring of more than 500 protein copies whose architecture was determined by combining cryo-electron tomography, cross-links, SAXS and structures of subcomplexes (Kim et al. 2018). Integrative models are archived in PDB-IHM (Section 2.5). Their reliability is assessed as for other models: by how precise the ensemble of solutions is, and, most convincingly, by whether they agree with data that were *not* used to build them. Fitting the same data repeatedly, or several data sets with correlated errors, increases apparent precision without adding evidence. The same principle applies when experimental data are used to check a predicted model (Section 9.8).

9.4.6 Experimental structures and predicted models

All experimental structures are models fitted to data, with errors that depend on the method and the resolution. Validation reports from the wwPDB compare each model with its data and with expected geometry (Burley et al. 2018). Since 2021, predicted structures have become as important as experimental ones for many purposes (Section 9.8). Experiments and prediction are now integrated: predicted models are used to phase crystallographic data, to interpret cryo-EM maps and to guide experiments, while experimental structures remain the ground truth on which predictors are trained and evaluated.

See also: further explanation in Section D.9.3; related Wikipedia articles in Section E.9.3.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Protein structure, Protein secondary structure, Ramachandran plot, X-ray crystallography, R-factor (crystallography), Protein nuclear magnetic resonance spectroscopy and Cryogenic electron microscopy (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Structure: Evidence, Models, and Dynamics" (CC BY 4.0), used with permission.

9.5 Protein folding

Many proteins fold from a chain of amino acids into a well-defined three-dimensional structure. The folding problem has three parts: the physical question of what determines the native structure, the kinetic question of how proteins fold so fast, and the computational question of predicting the structure from the sequence (Dill and MacCallum 2012). The third is the subject of Section 9.8. This section covers the first two, and what happens when folding goes wrong.

9.5.1 Anfinsen's dogma

In the 1950s and 1960s Christian Anfinsen showed that ribonuclease A, unfolded by urea and with its disulfide bonds reduced, refolds spontaneously into its active structure when the denaturant is removed. He concluded that, for a small globular protein **in a given environment** (solvent, pH, temperature, and any required cofactors or partners), **the information needed to specify the native structure is contained in the amino acid sequence**, and that the native structure is the conformation of lowest free energy under those conditions, the *thermodynamic hypothesis* (Anfinsen 1973). He received the 1972 Nobel Prize in Chemistry. The hypothesis is not the claim that sequence alone always determines a single structure. Anfinsen's dogma is the justification for predicting structure from sequence, and it has important exceptions and qualifications: the structure depends on the environment (a membrane, bound metals, cofactors or partner proteins); some proteins need chaperones to avoid aggregation (although chaperones do not change the final structure); some are kinetically trapped in states that are not the global minimum; some switch between folds; and many proteins or regions are intrinsically disordered (Section 9.8 and Section 3.5).

Even a well-folded protein does not have one rigid structure. The **native state** is an ensemble of closely related conformations, with side chains, loops and sometimes whole domains in motion, and some proteins populate two or more distinct functional states. A crystal structure or a predicted model shows one representative of this ensemble, often the most populated state or one stabilised by crystal contacts or bound ligands. Accuracy measures for predictions (Section 9.2) should be read with this in mind; dynamics and ensembles are discussed in Section 9.11.

9.5.2 Free energy and marginal stability

The native state is stable because of a balance of large opposing contributions. The **hydrophobic effect**, the tendency of non-polar side chains to avoid water, drives the burial of the core. Hydrogen bonds, van der Waals packing and electrostatic interactions stabilise the folded state, while folding costs a large loss of conformational entropy of the chain. The net result is that most proteins are only **marginally stable**: the free energy of unfolding, ΔG, is typically only 5–15 kcal/mol, the equivalent of a few hydrogen bonds. A single mutation can change it by 1–3 kcal/mol, which is why many disease mutations act by destabilising proteins, and why stability constrains evolution (Section 3.5). Small single-domain proteins often fold in a cooperative **two-state** transition, with only the unfolded and the native states populated at equilibrium (Jackson and Fersht 1991).

9.5.3 Levinthal's paradox and folding funnels

In the late 1960s Cyrus Levinthal pointed out that if each residue of a 100-residue chain could adopt only three conformations, there would be $3^{100} \approx 10^{48}$ conformations. Sampling them at a rate of 10^{13} per second would take far longer than the age of the universe, yet real proteins fold in microseconds to seconds (Levinthal 1966). Folding therefore cannot be a random search. Levinthal suggested that folding follows specific **pathways** guided by local interactions.

The modern resolution of the paradox is the **energy landscape** theory (Bryngelson et al. 1995; Dill and Chan 1997). Proteins evolved to have **funnel-shaped** free-energy landscapes: as the chain forms more native contacts, its energy decreases, so most moves lead downhill towards the native state, and there are many parallel routes rather than a single pathway. The width of the funnel represents the conformational entropy, which decreases as the chain folds. Real landscapes are rugged, with local minima in which the chain can be temporarily trapped, which explains intermediates and misfolding. Natural sequences are thought to be *minimally frustrated*: their interactions are mostly consistent with the native structure, unlike random sequences.

Experiments with protein engineering, measuring how mutations affect folding rates (*φ-value analysis*), showed which parts of the structure are formed in the transition state. Molecular dynamics simulations on the special-purpose computer Anton finally made it possible to watch small proteins fold and unfold repeatedly at atomic detail (Shaw et al. 2010; Lindorff-Larsen et al. 2011) (Section 9.11). They showed that fast-folding proteins fold along a small number of routes, often starting with native secondary structure, consistent with the funnel picture.

9.5.4 Misfolding, amyloid and prions

The native state is not the only deep minimum. Partly folded or unfolded chains expose hydrophobic segments and can **aggregate**, and one ordered aggregate is accessible to almost any polypeptide: the **amyloid fibril** (Chiti and Dobson 2017). Amyloid has a **cross-β** structure, in which β-strands run perpendicular to the fibril axis and are hydrogen-bonded into sheets that run along it, giving the characteristic X-ray fibre diffraction spacings of about 4.7 Å between strands and about 10 Å between sheets. Crystal structures of short amyloid-forming peptides showed pairs of sheets whose side chains interlock in a dry **steric zipper** (Sawaya et al. 2007). Because the backbone hydrogen bonds are shared by all sequences, amyloid can be more stable than the native fold at high concentration, so the native state of many proteins is only kinetically protected, by chaperones, degradation and their sequences, against aggregation. Anfinsen's hypothesis concerns the folding of a single chain, not this concentration-dependent alternative.

Aggregation is usually nucleated by short **aggregation-prone regions** of five to ten residues that are hydrophobic, have a high β-strand propensity and a low net charge. In folded proteins they are mostly

buried, and they are often flanked by charged residues or prolines that act as *gatekeepers*. Predictors identify them from sequence: **TANGO** uses a statistical-mechanics model of competing conformations (Fernandez-Escamilla et al. 2004), and **Waltz** uses position-specific scoring matrices trained on amyloid-forming hexapeptides (Maurer-Stroh et al. 2010). They are used to explain disease mutations and to reduce aggregation in therapeutic and designed proteins.

Prions are proteins whose misfolded, aggregated form converts the normal form into more of itself. Stanley Prusiner proposed that the agent of scrapie, and later of Creutzfeldt–Jakob disease and bovine spongiform encephalopathy, is a protein, the prion protein PrP, without nucleic acid (Prusiner 1982) (Nobel Prize in Physiology or Medicine, 1997). Wickner proposed, from genetic evidence, that the yeast trait [URE3] is a prion form of the Ure2 protein (Wickner 1994); yeast prions such as [URE3] and [PSI+] are inherited through the cytoplasm, so protein conformation can carry heritable information. The same templated growth occurs, in a less infectious form, in the spread of aggregates between cells in neurodegenerative diseases.

Fibrils are not crystals, and their atomic structures were long unknown. Since 2017, **cryo-EM** has determined many of them at near-atomic resolution (Section 9.4): tau filaments from the brain of a patient with Alzheimer's disease (Fitzpatrick et al. 2017), amyloid-β(1–42) (Gremer et al. 2017) and α-synuclein, the protein of Parkinson's disease (Guerrero-Ferreira et al. 2018). Each layer of a fibril is one chain folded into a flat, often intricate, arrangement of β-strands, quite different from its native or disordered state. The same protein can form several **polymorphs**; different tauopathies have different tau folds, which supports the idea that structural strains underlie different diseases, and fibrils formed *in vitro* often differ from those extracted from patients. AlphaFold and similar predictors, trained mainly on globular proteins and predicting single chains or defined complexes, do not reliably predict these folds.

Not all amyloid is pathological. **Functional amyloids** include the curli fibres of *Escherichia coli* biofilms, which are assembled by a dedicated secretion and nucleation machinery (Chapman et al. 2002), fibres in mammalian melanosomes, and stores of peptide hormones in secretory granules (Sawaya et al. 2021). Their existence shows that cells can control amyloid formation, and that the cross-β fold is part of normal biology.

9.5.5 Folding and prediction

AlphaFold and related methods predict the native structure of many proteins very accurately without simulating folding at all (Section 9.8). They learn the mapping from sequence and evolutionary information to structure from the PDB. This largely solved the practical problem of predicting a static structure for most single-chain globular proteins, but not the physical question of how folding happens. The models do not give folding pathways, are less reliable for the effects of single mutations on stability, and predict one structure even for proteins that switch between several.

See also: further explanation in Section D.9.4; related Wikipedia articles in Section E.9.4.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Protein folding, Anfinsen's dogma, Levinthal's paradox, Folding funnel and Hydrophobic effect (CC BY-SA 4.0, retrieved 4 October 2026).

9.6 Structure comparison and evolution

Comparing structures is needed to evaluate predictions, to classify proteins, and to detect evolutionary relationships that are no longer visible in the sequence. Among homologues, structure usually changes much more slowly than sequence: proteins with less than 20% sequence identity, in the "midnight zone" of sequence comparison (1), often have the same fold and superimpose closely (Chothia and Lesk 1986). This is an empirical tendency, not a law. A few mutations can switch a protein to a different fold, and similar folds can arise independently by convergence, so structural similarity alone does not prove common ancestry. With these caveats, structural similarity is often the most sensitive evidence for distant homology.

9.6.1 Measures of similarity

The simplest measure is the **root-mean-square deviation** (RMSD) between equivalent atoms (usually Cα atoms) after optimal superposition,

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \lVert \mathbf{x}_i - \mathbf{y}_i \rVert^2},$$

where the optimal rotation is found analytically, for example with the Kabsch algorithm. RMSD has two problems. It requires a known correspondence between residues, and it is dominated by the largest deviations, so two structures with an identical core and one different loop or a domain movement get a large RMSD. It also depends on the length of the protein.

The **TM-score** solves the last two problems (Zhang and Skolnick 2004):

$$\text{TM-score} = \max \frac{1}{L}\sum_{i=1}^{N} \frac{1}{1 + (d_i/d_0(L))^2},$$

where d_i is the distance between the ith pair of aligned residues, L the length of the reference protein, and $d_0(L)$ a length-dependent scale chosen so that random structure pairs give similar scores regardless of size. The maximum is taken over superpositions. Each residue contributes between 0 and 1, so distant pairs contribute little instead of dominating. TM-scores above 0.5 indicate generally the same fold, and below about 0.17 random similarity. **GDT_TS** and **lDDT**, used in CASP, are described in Section 9.2.

9.6.2 Structural alignment

Structural alignment finds the residue correspondence that makes two structures as similar as possible, without using the sequence. Since the correspondence and the superposition depend on each other, the problem is hard, and all programs use heuristics (Kolodny et al. 2005):

- **DALI**, by Liisa Holm and Chris Sander, compares the matrices of intramolecular Cα–Cα distances of the two proteins. It looks for sub-matrices (fragment pairs) with similar distance patterns and assembles them into an alignment with Monte Carlo optimisation (Holm and Sander 1993). Because it compares internal distances, it needs no superposition and can align structures with different topologies of the same contacts. DALI was used to map the protein structure universe (Holm and Sander 1996) and is still widely used (Holm 2019).

- **TM-align** iterates between superposition and alignment: it starts from several initial alignments, computes the superposition that maximises the TM-score, and realigns the residues with dynamic programming using the distances after superposition (Zhang and Skolnick 2005).
- **Foldseek** made structure search as fast as sequence search (van Kempen et al. 2023). It describes the local tertiary environment of each residue with a letter from a learned 20-letter *3Di* structural alphabet, so that each structure becomes a string, which can be searched with the fast prefilters and alignment algorithms of MMseqs2 (Section 1.4). Foldseek is thousands of times faster than DALI or TM-align with similar sensitivity, and made it possible to search and cluster the hundreds of millions of predicted structures in the AlphaFold Database (Barrio-Hernandez et al. 2023).

9.6.3 What structure comparison has shown

Comparisons of the growing number of structures showed that the number of distinct folds is limited. Chothia estimated in 1992 that there might be about 1,000 folds (Chothia 1992), and the structure classifications show that a few hundred *superfolds*, such as the TIM barrel, the Rossmann fold and the immunoglobulin fold, account for most domains. As more structures were determined, new folds became rarer. The AlphaFold Database, covering almost all known proteins, confirmed this. Clustering its 214 million predicted structures with Foldseek gave 2.3 million non-singleton clusters. About 31% of the clusters had no annotation and may represent previously undescribed structures, but together they contain only 4% of the proteins: most proteins belong to known families. The comparison also revealed remote relationships, for example between human immune-related proteins and prokaryotic ones (Barrio-Hernandez et al. 2023). Structure comparison has also shown that proteins with the same function can have different folds (convergent evolution), and that the same fold can carry very different functions, which limits how far function can be transferred by structural similarity (Chapter 10).

See also: further explanation in Section D.9.5; related Wikipedia articles in Section E.9.5.

9.7 Protein structure databases

The databases of structures are described in detail in Section 2.5, and the domain classifications in Section 1.9. In brief:

- The **Protein Data Bank** (PDB), founded in 1971, is the single archive of experimental structures, about 260,000 entries in 2026 (Berman et al. 2000; Burley et al. 2018).
- **SCOP** (1995) classifies domains manually by class, fold, superfamily and family, with superfamilies defined by evidence of common ancestry (Murzin et al. 1995; Andreeva et al. 2019). Its four structural classes go back to

  2.

- **CATH** classifies domains by Class, Architecture, Topology (fold) and Homologous superfamily, with more automation (Sillitoe et al. 2020).
- **ECOD** emphasises evolutionary relationships, with a level of "possible homology" above the superfamily, and is updated weekly (Cheng et al. 2014).
- **TED** (The Encyclopedia of Domains) applied automatic domain parsing and classification to the AlphaFold Database, identifying hundreds of millions of domains in predicted structures (Lau et al. 2024).
- The **AlphaFold Protein Structure Database** contains predicted structures for more than 200 million proteins, almost all of UniProt (Varadi et al. 2023; Bertoni et al. 2025).
- The **ESM Metagenomic Atlas** contains more than 600 million structures of metagenomic proteins predicted with ESMFold (Lin et al. 2023) (Section 9.9).

See also: further explanation in Section D.9.6; related Wikipedia articles in Section E.9.6.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Root-mean-square deviation of atomic positions, Template modeling score, Structural alignment, Structural Classification of Proteins database and CATH database (CC BY-SA 4.0, retrieved 4 October 2026).

9.8 Protein structure prediction

Predicting the structure of a protein from its sequence was one of the grand challenges of computational biology for fifty years. Anfinsen's dogma (Section 9.5) said that it should be possible; the size of the conformational space and the inaccuracy of energy functions made it hard. The history of the field is a sequence of ideas, each of which used more of the information hidden in evolution, until deep learning combined them in 2020. Progress was measured throughout by CASP (Section 9.2).

9.8.1 Secondary structure prediction

The first methods predicted only secondary structure, the state (helix, strand or coil) of each residue.

- **Chou–Fasman** (1974) used the propensity of each amino acid to occur in helices and strands, computed from the few structures then known, with rules for nucleating and extending elements (Chou and Fasman 1974a, 1974b).
- **GOR** (1978) used information theory to include the influence of neighbouring residues in a window of 17 residues (Garnier et al. 1978).

Both reached about 60% accuracy (Q3, the percentage of residues correctly predicted in three states), little better than chance because local sequence only partly determines local structure. The breakthrough came from evolution. **PHD**, by Burkhard Rost and Chris Sander, fed **multiple sequence alignments** of homologous proteins into a neural network, raising accuracy above 70% (Rost and Sander 1993): a family of sequences shows which positions are conserved, hydrophobic or variable, which the single sequence cannot. **PSIPRED** used PSI-BLAST profiles and reached about 77–80% (Jones 1999). Modern deep-learning methods reach about 85%, close to the limit set by the ambiguity of assigning secondary structure in experimental structures.

9.8.2 Assigning secondary structure from coordinates

To train and test predictors, the "true" secondary structure must first be assigned from experimental coordinates, and this is less objective than it sounds. **DSSP** (Kabsch and Sander, 1983), the standard program, finds backbone hydrogen bonds with a simple electrostatic model: partial charges on the C=O and N–H groups give an energy, and a hydrogen bond is counted if it is below -0.5 kcal/mol (Kabsch and Sander 1983). Secondary structure is then defined by *patterns* of hydrogen bonds: repeated $i \to i+4$ turns form an α-helix, and ladders of bridges between strands form a β-sheet. DSSP assigns eight states:

- **H**: α-helix; **G**: 3_{10}-helix; **I**: π-helix;
- **E**: strand in a β-ladder; **B**: isolated β-bridge;
- **T**: hydrogen-bonded turn; **S**: bend (a geometric criterion only);
- and blank (often written C or "–"): none of these, loosely called coil.

For prediction, the eight states are reduced to three. A common reduction is H, G, I → helix; E, B → strand; everything else → coil, but some benchmarks map only H and E, and the choice changes Q3 by a few percentage points, so the reduction must be stated when methods are compared. **STRIDE** combines hydrogen-bond energies with backbone torsion angles, with parameters tuned to agree with the assignments made by crystallographers (Frishman and Argos 1995). DSSP and STRIDE agree for most residues but differ at the ends of helices and strands, for 3_{10} and π helices, for irregular or bulged strands, and in low-resolution structures; NMR models of one protein can differ among themselves. This disagreement is one reason why Q3 cannot reach 100%.

Q3 also has a weakness as a measure: it counts residues, not elements. A prediction that shifts the ends of every helix by one residue, and one that breaks a helix into fragments or misses a short strand completely, can have the same Q3. The **segment overlap score** (SOV) compares predicted and observed *segments* and tolerates small shifts at their boundaries, so it rewards predictions that get the number, type and approximate position of the elements right (Rost et al. 1994; Zemla et al. 1999). Q3 and SOV are usually reported together.

9.8.3 Homology modelling

If a protein of known structure is homologous to the query, the structure of the query can be modelled on it, since homologous proteins usually have similar structures (Section 9.6). **Homology modelling** (comparative modelling) has four steps:

1. **Template search**: finding related structures, with sequence or profile searches against the PDB (Chapter 1).
2. **Alignment** of query and template, the most critical step; errors here cannot be corrected later.
3. **Model building**: copying the coordinates of aligned regions, and modelling loops and side chains. **MODELLER** builds the model by satisfying spatial restraints derived from the template, such as distances and angles, together with stereochemical rules (Šali and Blundell 1993).
4. **Evaluation** of model quality.

The accuracy of the model depends mainly on the sequence identity to the template: above about 50%, models are close to medium-resolution experimental structures, while below 30% alignment errors dominate. Tools such as SWISS-MODEL made homology modelling routine.

9.8.4 Loops and side chains

After the aligned core has been copied, two parts of a homology model remain: loops and side chains. **Loops**, the insertions relative to the template and the regions where the template differs, are usually the least accurate part of a model, and accuracy falls with loop length. *Database* methods search known structures for fragments whose ends fit the geometry of the two anchor residues and whose sequences are similar; they work well for short loops. *Ab initio* (conformational search) methods sample loop conformations and score them with an energy function, as MODELLER does (Fiser et al. 2000). A loop must close exactly between its anchors, which makes random sampling inefficient. **Kinematic closure** (KIC) in Rosetta, borrowed from robotics, samples the torsions of most loop residues freely and solves analytically for six torsions that close the chain, and reconstructed many 12-residue loops with sub-ångström accuracy (Mandell et al. 2009).

Side chains adopt a limited number of preferred combinations of their torsion angles $\chi_1, \chi_2, \ldots$, called **rotamers**, mostly with staggered torsions. **Rotamer libraries** list them with their frequencies (Ponder and Richards 1987). Dunbrack and Karplus showed that the frequencies depend on the backbone torsions ϕ and ψ, and the **backbone-dependent rotamer library** (Dunbrack and Karplus 1993), now a smoothed version derived from high-resolution structures (Shapovalov and Dunbrack 2011), is used by most

modelling and design programs. Placing side chains then becomes a combinatorial problem: choose one rotamer per residue to minimise an energy with terms for each rotamer and each pair of rotamers. **Dead-end elimination** removes rotamers that cannot be part of the global minimum (Desmet et al. 1992); **SCWRL4** splits the graph of interacting residues into small clusters that can be solved exactly (Krivov et al. 2009); and the Rosetta **packer** uses Monte Carlo with simulated annealing over rotamers (Section 9.11). The packer is the same algorithm used for design, where the rotamers of all allowed amino acids compete at each position (Section 9.12). Buried side chains are placed much more accurately than exposed ones, and accuracy drops when the backbone itself is inaccurate.

Rosetta, which began as a fragment-assembly program (below), grew into a software suite that shares one energy function (Alford et al. 2017) across loop modelling, packing, refinement, docking and design (Leman et al. 2020). MODELLER instead optimises all restraints of the model together, and SWISS-MODEL builds models from templates with its own loop and side-chain modelling. Deep-learning predictors removed these separate steps. AlphaFold2 predicts side-chain torsion angles in its structure module, and AlphaFold3 generates all atoms with diffusion, so loops and side chains are predicted together with the rest of the structure and are generally reliable where pLDDT is high. Rotamer libraries and packers remain essential in design, for modelling mutations and for refining models against experimental data.

9.8.5 Fold recognition and threading

Many proteins have a known fold but no detectable sequence similarity to it. **Fold recognition** asks which known fold a sequence is most compatible with. Bowie, Lüthy and Eisenberg described each structure as a **3D profile**, the environment of each position (buried or exposed, in which secondary structure), and scored sequences against it (Bowie et al. 1991). David Jones and colleagues introduced **threading**: the query sequence is placed onto the backbone of each template and scored with statistical pair potentials derived from known structures (Jones et al. 1992). Later methods combined sequence profiles with predicted secondary structure and solvent accessibility (Fischer and Eisenberg 1996), and **profile–profile** comparison (HHpred and similar methods) became the most sensitive way to detect templates (Section 1.6).

9.8.6 De novo prediction and fragment assembly

For proteins with no template, structures must be predicted from scratch, by searching conformational space for low-energy structures. Two problems make this hard: the search space is enormous (Section 9.5), and the energy functions are not accurate enough to recognise the native state reliably. **Fragment assembly** reduced the search space: short fragments (three and nine residues) from known structures, chosen by local sequence similarity, are combined by Monte Carlo search, and scored with knowledge-based energy functions. David Baker's **Rosetta** (Simons et al. 1997) and David Jones's **FRAGFOLD** (Jones 2001) produced the first successful de novo predictions of small proteins in CASP. With extensive sampling and all-atom refinement, Rosetta could reach atomic accuracy for some small proteins (Bradley et al. 2005). In general, however, de novo methods worked only for proteins smaller than about 100 residues.

9.8.7 Coevolution

The idea that would eventually lead to the breakthrough was old. Residues in contact in a structure must remain compatible: if one mutates, its partner often mutates in a compensating way. Such **correlated mutations** should be visible in multiple sequence alignments (Göbel et al. 1994). For two decades this

had little practical effect, because correlations are also transmitted *indirectly*: if A contacts B and B contacts C, A and C appear correlated without being in contact.

The solution was to fit a **global statistical model** of the whole alignment, a Potts model (a Markov random field) with fields for each position and couplings for each pair, so that direct couplings can be separated from indirect correlations. This **direct-coupling analysis** (DCA) (Weigt et al. 2009; Morcos et al. 2011), and fast approximations such as pseudo-likelihood methods and sparse inverse covariance estimation (PSICOV) (Jones et al. 2011), predicted contacts accurate enough to fold proteins from sequence alone (Marks et al. 2011; Marks et al. 2012). The method requires many sequences, typically more than a few times the protein length. Metagenomic sequences extended it to many more families (Ovchinnikov et al. 2017).

Contacts form patterns, such as the parallel diagonals of β-sheets, so the next step was to predict contact maps with machine learning that recognised these patterns (Skwark et al. 2014), and then with deep convolutional networks that treated the coevolution matrix as an image (Wang et al. 2017). In CASP13 (2018), networks predicting full distance distributions rather than binary contacts, combined with gradient-descent optimisation of the structure, gave a large improvement; DeepMind's first AlphaFold was the best of these (Senior et al. 2020).

9.8.8 AlphaFold2

At CASP14 (2020), **AlphaFold2** produced models comparable to experimental structures for most targets (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021). Its main components are:

- **Inputs**: a multiple sequence alignment of the query, built with sensitive searches of large sequence databases including metagenomes, and optionally structural templates. Even when no template is supplied, the trained network has encoded structural knowledge from the PDB, so "template-free" should be read as "no explicit template used", not as prediction without prior structural knowledge.
- **The Evoformer**: a deep network of 48 blocks that maintains two representations, an MSA representation (sequences × residues) and a pair representation (residues × residues). Attention operates along rows and columns of the MSA, information flows between the two representations, and *triangle updates* on the pair representation enforce geometric consistency, analogous to the triangle inequality of distances.
- **The structure module** converts the pair and single representations into 3D coordinates. Each residue is treated as a rigid body (a frame), positioned with an attention mechanism that is invariant to global rotation and translation (*invariant point attention*), and side-chain torsion angles are predicted.
- **Recycling**: the output is fed back through the network several times.
- **Training** on structures in the PDB, followed by *self-distillation* on high-confidence predictions of unlabelled sequences, with a loss based on frame-aligned errors and auxiliary losses on distances and on the masked MSA.

AlphaFold2 also predicts its own accuracy. These measures are the network's estimates, not measured accuracy, and each answers a different question:

- **pLDDT** is a predicted lDDT score for each residue: the model's confidence in the *local* structure around that residue. It is not the probability that the region is folded, and says nothing about the relative placement of domains. Low pLDDT often indicates intrinsically disordered regions (see below).

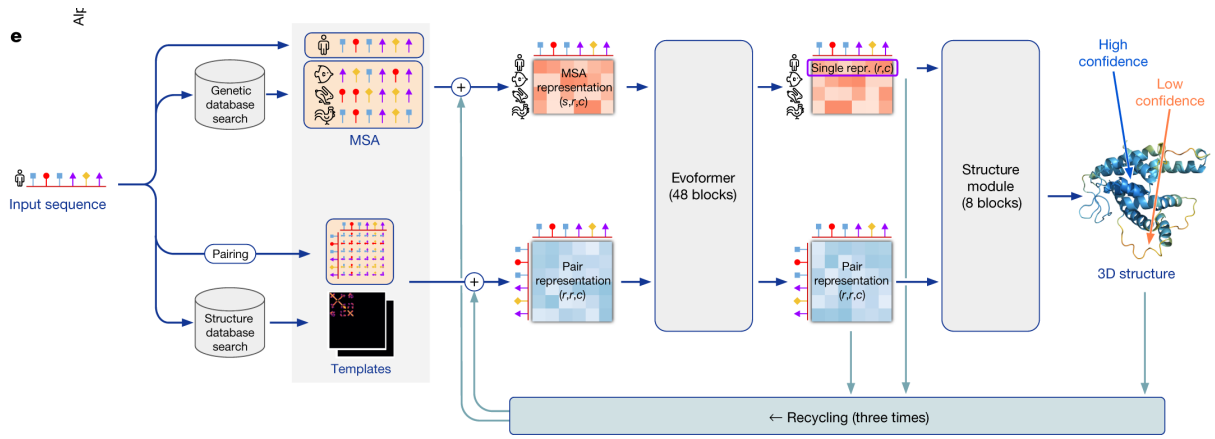


Figure 9.2: The AlphaFold2 architecture. The input sequence is used to search genetic databases for an MSA and structure databases for templates; the MSA and pair representations are updated by the 48 Evoformer blocks; the structure module (8 blocks) turns the single and pair representations into 3D coordinates with per-residue confidence; and the outputs are recycled three times. Figure 1e from Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. (2021) (CC BY 4.0), cropped by Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

- The **predicted aligned error** (PAE) is the expected error in the position of residue j when the predicted and true structures are aligned on residue i. It shows whether the relative positions of domains, or of chains in a complex, are reliable.
- **pTM** and, for complexes, **ipTM** are global scores derived from the PAE for the whole model and for the interfaces between chains (Section 11.4). Confident individual chains (high pLDDT) do not imply a correct interface.

Using these measures correctly is essential: a predicted structure is a hypothesis with an estimated reliability, not a measurement.

The code was released in 2021, and DeepMind and EMBL-EBI used it to predict structures for almost all known proteins (Section 9.7). Demis Hassabis and John Jumper shared the 2024 Nobel Prize in Chemistry with David Baker. **RoseTTAFold**, from Baker's group, used a three-track network (sequence, distance and 3D coordinates) and reached similar accuracy (Baek et al. 2021). **ColabFold** made prediction fast and accessible by replacing the slow alignment step with MMseqs2 searches (Mirdita et al. 2022), and **OpenFold** retrained AlphaFold2 from scratch, showing how it learns (Ahdritz et al. 2024).

9.8.9 Complexes and co-folding: AlphaFold3 and successors

Soon after its release, AlphaFold2 was found to predict the structures of protein complexes when sequences of several chains were joined, and **AlphaFold-Multimer** was trained specifically for complexes (Evans et al. 2021; Bryant, Pozzati, and Elofsson 2022) (Chapter 11).

AlphaFold3 generalised the approach to almost all molecules in the PDB: proteins, DNA, RNA, small-molecule ligands, ions and modified residues (Abramson et al. 2024). It replaced the Evoformer with a simpler *Pairformer*, which uses the MSA less, and the structure module with a **diffusion model** that generates all-atom coordinates directly, starting from noise and denoising them step by step. On several retrospective benchmarks, this *co-folding* of all components predicted protein–ligand poses substantially better than classical docking methods (Section 9.11). Classical docking still has advantages

in some workflows, for example for explicit sampling, scoring very many ligands or adding constraints, so the question is not settled. Because AlphaFold3 was initially released without code, several groups reimplemented and extended the approach with open code: **Chai-1** (Boitreaud et al. 2024), **Boltz-1** (Wohlwend et al. 2024) and **Protenix** (Chen et al. 2025) report AlphaFold3-level accuracy, and **Boltz-2** also predicts binding affinities (Passaro et al. 2025) (all first released as preprints).

Three systems released in 2026 show different ways of moving beyond AlphaFold3. They are very recent, and their status (as of October 2026) differs: a preprint, a company technical report and a software repository. None of them is a peer-reviewed publication.

- **ESMFold2** (Candido et al., bioRxiv preprint, June 2026), from the ESM team, joins the large protein language model ESM C (Section 9.9) to an AlphaFold3-style diffusion structure module. Unlike the original ESMFold, it predicts complexes of proteins, nucleic acids and ligands, from single sequences or with an optional alignment. The authors report that it outperforms established methods for complex prediction, including antibody–antigen interactions, and that it can be used to design binders with nanomolar affinity (Candido et al. 2026). According to the authors, the model weights are openly released.
- **IsoDDE**, the drug design engine of Isomorphic Labs, DeepMind's drug-discovery sister company, is proprietary. Its technical report (February 2026, not peer reviewed) claims more than twice AlphaFold3's accuracy for the protein–ligand complexes least similar to the training data, 2.3 times AlphaFold3's success rate at high accuracy (DockQ > 0.8) for new antibody–antigen complexes, affinity predictions that can compete with physics-based free-energy calculations, and the identification of binding pockets from sequence (Isomorphic Labs Team 2026). There is no code, no model and no independent test, so these are claims, not results. Notice that the claimed gains are largest for examples unlike the training data, where AlphaFold3 is weakest.
- **OpenDDE** (a GitHub software repository, 2026) is not a new model and not a publication, but a community project that chains existing open tools behind one interface: AlphaFold3 for complexes, P2Rank for pockets, Boltz-2 for affinity, ImmuneBuilder for antibodies, and ChEMBL and RDKit for known ligands (OpenDDE contributors 2026). It has no benchmarks of its own. It shows what a drug design "engine" is: a pipeline of separate models, each of which can, and should, be evaluated on its own.

The contrast matters for science. Open tools such as ESMFold2, Boltz-2 and OpenFold can be tested by anyone on data they have never seen, while a closed system can only be judged by what its owners choose to report. Claims of accuracy on "new" targets should always be checked for overlap with the training data (4).

9.8.10 Intrinsically disordered proteins

Not all proteins fit the picture of one sequence, one stable structure. **Intrinsically disordered proteins and regions** (IDPs and IDRs) have no stable tertiary structure on their own and exist as dynamic ensembles of conformations. They are common in eukaryotes, especially in signalling and transcription regulation, and they evolve differently from folded domains (Section 3.5) (Lee et al. 2014; Brown et al. 2011; Light et al. 2013). Many fold, partly or fully, only when they bind a partner (*folding upon binding*) or under particular conditions, for example after phosphorylation: their structure is *conditional*.

Disorder can be predicted from sequence, because disordered regions have a characteristic composition: few hydrophobic and aromatic residues, many charged and polar residues, prolines and glycines, and often low sequence complexity. Dedicated **disorder predictors** exploit this, from IUPred, which estimates whether a segment can form enough favourable inter-residue contacts to fold (Erdős et al. 2021), to neural networks using alignments or language-model embeddings; they are assessed in blind community experiments (CAID) in the spirit of CASP (Necci et al. 2021). AlphaFold2's **low pLDDT**

(below about 50) is also a competitive disorder predictor, which is why models of eukaryotic proteins in the AlphaFold Database often contain long, extended, low-confidence regions (Tunyasuvunakool et al. 2021). These regions should not be read as structures. Two caveats apply. Low pLDDT means only that no confident structure was predicted, which may reflect disorder or simply a lack of information. Conversely, conditionally folded regions are sometimes predicted with high confidence in their bound conformation, which is then a hypothesis about the bound state, not about the free protein (Alderson et al. 2023). Neither a single model nor a confidence score describes the ensemble of a disordered region; that requires simulation and experiments such as NMR (Section 9.11).

9.8.11 Limitations and current questions

Prediction is not finished. Current methods:

- predict a single structure, while many proteins have several functional conformations; sampling alternative states, for example by clustering the MSA, is an active research area (Wayment-Steele et al. 2023), but such samples are plausible alternative states, not a thermodynamic ensemble with equilibrium populations (Section 9.11);
- depend on the MSA, and are less accurate for orphan proteins, rapidly evolving proteins and designed or mutated sequences, and do not reliably predict the effects of point mutations;
- are much less reliable for antibody–antigen and other complexes without coevolution between the partners, for RNA, and for large flexible assemblies;
- can produce confident but wrong predictions ("hallucinations"), particularly for disordered regions in diffusion-based models, and ligand poses that are physically implausible (Buttenschoen et al. 2024);
- by default predict the bare polypeptide: cofactors, metal ions, glycans and post-translational modifications are absent unless they are given explicitly (AlphaFold3 and similar models accept some of them as input), so the model may not represent the biologically relevant molecular state, for example a metal-binding site without its metal, or a protein whose conformation depends on phosphorylation;
- learn from the PDB, so they reflect its biases, and they say nothing about the folding process (Section 9.5).

9.8.12 Checking a predicted model

Before conclusions are drawn from a predicted structure, check:

- **Confidence**: pLDDT for local structure, PAE for the relative placement of domains and chains, and ipTM or a similar score for interfaces; read each for what it measures.
- **Stereochemistry**: clashes, bond geometry, Ramachandran outliers and, for ligands, chemically valid poses, with the same validation tools as for experimental structures (Burley et al. 2018; Buttenschoen et al. 2024).
- **Leakage**: whether the protein, or a close homologue, was among the templates or training structures; if so, high accuracy is expected rather than impressive (4).
- **Oligomeric state**: whether the protein acts as a monomer or in a complex; a single subunit predicted alone may have exposed interface surfaces or unstructured segments (Chapter 11).
- **Ligands, cofactors and modifications**: whether the biologically relevant state includes metals, cofactors, glycans or modifications that were not modelled.
- **Consistency with experiments**: agreement with available data such as mutagenesis, crosslinking, small-angle scattering, cryo-EM maps or known functional residues.

See also: further explanation in Section D.9.7; related Wikipedia articles in Section E.9.7.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Protein structure prediction, Protein secondary structure, Homology modeling, Threading (protein sequence), De novo protein structure prediction, Direct coupling analysis, Intrinsically disordered proteins and AlphaFold (CC BY-SA 4.0, retrieved 4 October 2026).

9.9 Protein language models and structure

Protein language models (pLMs) are trained on hundreds of millions of protein sequences to predict masked or next amino acids, in the same way as language models for text. Their architecture and training are described in Section 1.10. This section covers what they contribute to structure prediction. The key observation is that, to predict missing residues well, a model must learn which positions co-vary, which is the coevolution signal of Section 9.8, and this information appears in the attention maps and embeddings without any structure ever being shown to the model.

9.9.1 ESMFold and the ESM Metagenomic Atlas

The **ESM-2** models from Meta AI, with up to 15 billion parameters, showed that structural information in the embeddings increases with model size. **ESMFold** adds a folding trunk and an AlphaFold-like structure module on top of ESM-2, and predicts structures from a **single sequence**, without an MSA (Lin et al. 2023). It is less accurate than AlphaFold2 for proteins with few homologues, but an order of magnitude or more faster, since the slow alignment step is not needed. Its accuracy correlates with how well the language model "understands" the sequence (its perplexity), which suggests that it relies on evolutionary information stored in the weights rather than on physics.

This speed made it possible to predict the structures of more than 600 million proteins from metagenomes, the **ESM Metagenomic Atlas** (Lin et al. 2023). About a third of these predictions had high confidence, and many had no similarity to known structures, giving a first view of the structural diversity of uncultured organisms (Section 5.7). The Atlas can be searched with Foldseek (Section 9.6).

9.9.2 ESM C and ESM3

After Meta's protein team became the company EvolutionaryScale, it released two successors:

- **ESM C** (ESM Cambrian, 2024) is a family of representation models with 300 million, 600 million and 6 billion parameters, trained on UniRef and metagenomic sequences, intended as a more efficient drop-in replacement for ESM-2: the 300-million-parameter model matches the 650-million parameter ESM-2 (EvolutionaryScale 2024). It was described in a technical blog post rather than a peer-reviewed paper. **ESMFold2** combines ESM C with a diffusion structure module to predict complexes (Section 9.8; a 2026 preprint) (Candido et al. 2026).
- **ESM3** is a generative model that reasons jointly over sequence, structure (encoded as discrete tokens) and function annotations (Hayes et al. 2025). It can be prompted with any combination of these, for example with a partial structure and a function keyword, and generates the rest. ESM3 was used to design a new fluorescent protein distant from all known ones (Section 9.12).

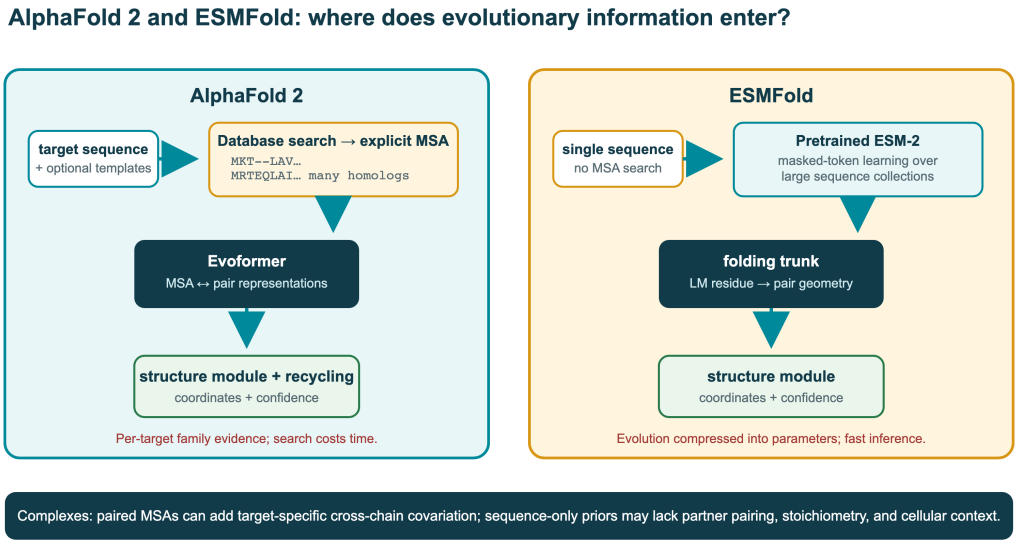


Figure 9.3: Where evolutionary information enters AlphaFold2 and ESMFold. AlphaFold2 searches databases for an explicit MSA of the target family, which the Evoformer combines with a pair representation; ESMFold takes a single sequence and uses the ESM-2 language model, in which information from millions of sequences has been compressed into the parameters during pre-training. For complexes, paired MSAs can add coevolution between chains, which a single-sequence model lacks. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

9.9.3 Structure-aware language models: ProstT5

A complementary idea is to teach language models structure explicitly. Foldseek's **3Di alphabet** (Section 9.6) turns each structure into a string, so sequence and structure can be treated as two languages. **ProstT5** is a protein language model fine-tuned to translate between amino acid sequences and 3Di strings (Heinzinger et al. 2024). Predicting 3Di sequences from amino acid sequences allows structure-based search with Foldseek even for proteins with no predicted structure, thousands of times faster than predicting the structure first. The reverse direction, from 3Di to amino acids, is a form of inverse folding, related to protein design (Section 9.12).

9.9.4 Language models versus alignments

Language models remove the need to build an MSA for each query, which is often the slowest step in prediction, and work for single sequences such as designed proteins and antibodies. MSA-based methods remain more accurate when many homologues are available. Whether the models learn general principles of protein structure, or mainly memorise the evolutionary information of protein families seen during training, is debated, and affects how well they generalise to sequences unlike those in nature (Chapter 12).

See also: further explanation in Section D.9.8; related Wikipedia articles in Section E.9.8.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Large language model and AlphaFold (CC BY-SA 4.0, retrieved 4 October 2026); ESM C is described in an EvolutionaryScale blog post.

9.10 Membrane proteins

Membrane proteins make up about a quarter of the proteins in most genomes and about half of all drug targets, but for a long time they were a small fraction of the PDB, because they are hard to express, purify and crystallise (Elofsson and Heijne 2007). Their structures are strongly constrained by the membrane, which made their topology one of the first aspects of structure that could be predicted reliably. There are two structural classes: **α-helical** proteins, found in all cell membranes, and **β-barrel** proteins, found in the outer membranes of Gram-negative bacteria, mitochondria and chloroplasts.

9.10.1 α-helical membrane proteins

The hydrophobic core of a lipid bilayer is about 30 Å thick. A polypeptide crossing it must satisfy the hydrogen bonds of its backbone, which an α-helix does internally. A **transmembrane helix** is therefore a stretch of about 20 mostly hydrophobic residues. Candidate segments can be found with **hydrophobicity plots**, which average a hydrophobicity scale over a sliding window. A hydropathy plot alone is not a sufficient predictor, however: signal peptides and some soluble helices are also hydrophobic; some transmembrane helices are only marginally hydrophobic and insert only together with neighbouring helices; and **re-entrant** loops or helices dip into the membrane and leave it on the same side.

The **topology** of a membrane protein, which segments cross the membrane and which side the loops are on, is largely governed by a few rules. The most important is the **positive-inside rule**, discovered by Gunnar von Heijne in Stockholm: loops on the cytoplasmic side are enriched in the positively charged residues lysine and arginine (Heijne 1992). The first topology predictors combined hydrophobicity with the positive-inside rule. **TMHMM** formulated the problem as a hidden Markov model, with states for helix cores, helix caps, cytoplasmic and non-cytoplasmic loops, whose grammar enforces the alternation of inside and outside loops (Krogh et al. 2001). It predicts transmembrane helices with about 97% accuracy and full topologies correctly for most proteins. Predictors must also distinguish transmembrane helices from **signal peptides**, which are also hydrophobic but are cleaved off (Chapter 10). **DeepTMHMM** replaced the hand-crafted features with embeddings from a protein language model, combined with a constrained output layer that keeps the grammar of the HMM, and predicts α-helical and β-barrel proteins and signal peptides together (Hallgren et al. 2022).

The structure of membrane proteins is now mostly predicted with general methods such as AlphaFold, which perform well on them. Topology predictors remain useful as fast tools for whole proteomes and to orient predicted structures in the membrane.

9.10.2 β-barrel membrane proteins

Outer membrane proteins of Gram-negative bacteria have a completely different architecture: an antiparallel β-sheet of 8 to 26 strands closed into a **barrel**, with hydrophobic residues facing the lipids and polar residues lining the interior (Wimley 2003). They include porins, transporters and enzymes. In each membrane-crossing strand, residues alternate between pointing into the membrane and into the barrel, so the strands are shorter (about 9–11 residues) and less hydrophobic on average than transmembrane helices, and hydrophobicity plots do not find them. β-barrel predictors therefore use the alternating pattern, the even number of strands and the fact that loops alternate between the periplasm and the outside. PRED-TMBB used a hidden Markov model (Bagos et al. 2004), and BOCTOPUS added the dyad-repeat pattern of strands (Hayat et al. 2016). DeepTMHMM predicts both classes with a single model (Hallgren et al. 2022).

See also: further explanation in Section D.9.9; related Wikipedia articles in Section E.9.9.

Sources

Parts of this section are adapted from the Wikipedia articles Membrane protein, Transmembrane domain, Membrane topology and Beta barrel (CC BY-SA 4.0, retrieved 4 October 2026).

9.11 Dynamics and docking

Proteins are not static. They vibrate, side chains rotate, loops and domains move, and many proteins switch between conformations as part of their function. They also bind other molecules. This section describes physics-based simulation of protein motion and the prediction of how proteins bind small molecules.

9.11.1 Molecular dynamics

Molecular dynamics (MD) simulates the motion of every atom by integrating Newton's equations of motion, $m_i \ddot{\mathbf{x}}_i = -\nabla_i U(\mathbf{x})$, in time steps of about 2 femtoseconds. The forces come from a **force field**, an empirical potential energy function $U(\mathbf{x})$ with terms for bond lengths, bond angles and torsions, and non-bonded van der Waals and electrostatic interactions, with parameters fitted to quantum-mechanical calculations and experimental data. Water is included explicitly as thousands of molecules, or approximated as a continuum.

The first simulation of a protein, the bovine pancreatic trypsin inhibitor for 9 picoseconds in vacuum, was published in 1977 (McCammon et al. 1977). Levitt and Warshel had already simulated protein folding with a simplified model (Levitt and Warshel 1975), and Warshel and Levitt introduced combined quantum mechanical/molecular mechanical (QM/MM) methods for enzyme reactions (Warshel and Levitt 1976). Martin Karplus, Michael Levitt and Arieh Warshel received the 2013 Nobel Prize in Chemistry for these multiscale models (Karplus and McCammon 2002).

The main limitation of MD is **time scale**. Biologically important events, such as folding, large conformational changes and ligand unbinding, take microseconds to seconds, while a simulation needs about a billion steps per microsecond. Advances in hardware, from graphics processors to the special-purpose supercomputer **Anton**, reached millisecond simulations and allowed small proteins to be folded at atomic detail (Shaw et al. 2010; Lindorff-Larsen et al. 2011). *Enhanced sampling* methods, such as replica exchange and metadynamics, accelerate the crossing of energy barriers. The other limitation is **accuracy of the force field**: simulations are only as good as the energy function, and systematic comparisons with experimental data have driven steady improvements (Lindorff-Larsen et al. 2012). MD is used to study mechanisms, to refine structures, to calculate free energies of binding and to sample the conformational ensembles that static structure prediction misses.

9.11.2 Normal modes and elastic network models

Large-scale, slow motions of proteins, such as hinge bending between domains, can be studied without full MD. **Normal mode analysis** approximates the energy around a minimum as a harmonic function and computes the vibrational modes of the structure. Tirion showed that a very simple model, in which all atoms within a cutoff distance are connected by identical springs, reproduces the slowest modes well (Tirion 1996). **Elastic network models**, such as the Gaussian network model (Bahar et al. 1997) and the anisotropic network model (Atilgan et al. 2001), use only the Cα atoms. The few lowest-frequency modes often describe functional motions, and the models predict which residues are flexible, in agreement with crystallographic temperature factors, at a tiny fraction of the cost of MD.

9.11.3 Structures versus ensembles

Predicting a structure and predicting a conformational ensemble are different problems. A structure predictor returns one or a few static models. A protein in solution populates an ensemble of conformations, with relative populations set by their free energies (Section 9.5). Function often depends on the ensemble: in **allostery**, an effector shifts the populations between states; and a ligand may bind a conformation that is rare without it (*conformational selection*) or induce a change on binding (*induced fit*). Running AlphaFold with different random seeds, with subsampled or clustered MSAs or with dropout gives structural diversity, which can reveal alternative states (Wayment-Steele et al. 2023). It does not, however, sample a thermodynamic ensemble, and the frequencies of the models are not equilibrium populations. Ensembles and populations come from MD and enhanced sampling, from experiments such as NMR and cryo-EM classification (Section 9.4), or from methods that combine learned models with physics, an active research area.

9.11.4 Sampling, convergence and computational cost

At equilibrium, the population of a conformational state i with free energy G_i follows the **Boltzmann distribution**,

$$p_i = \frac{e^{-G_i/RT}}{\sum_j e^{-G_j/RT}},$$

so a free-energy difference of about 1.4 kcal/mol at 25 °C corresponds to a tenfold difference in population (Section 9.3). Simulation methods aim to sample from this distribution. **Monte Carlo** methods do so without following time: a random move that changes the energy from U to U' is accepted with the Metropolis probability

$$P_{\text{accept}} = \min\left\{1, e^{-(U'-U)/k_B T}\right\},$$

so downhill moves are always accepted and uphill moves sometimes (Metropolis et al. 1953). If the moves can reach all states, the samples follow the Boltzmann distribution of the energy function. The same algorithm is used for Bayesian phylogenetics (Section 4.3), for fragment assembly in Rosetta (Section 9.8) and for sequence design (Section 9.12); lowering the temperature gradually (*simulated annealing*) turns it into a search for low-energy states.

Sampling is the main source of error in both MD and Monte Carlo. Successive frames of a trajectory are strongly correlated, so a long simulation that stays in one basin contains less information than its length suggests. Convergence is tested by running independent simulations from different starting structures, comparing averages over separate blocks of the trajectory, and comparing computed observables with experiments such as NMR or SAXS. A longer simulation cannot correct a wrong starting point, for example a missing ligand or a wrong protonation state.

The cost of each method grows differently with the size of the system. In MD, short-range forces are computed with neighbour lists at a cost roughly proportional to the number of atoms N, and long-range electrostatics with particle-mesh Ewald methods at a cost close to $N \log N$; with time steps of 2 fs, one microsecond needs some 5×10^8 steps. In AlphaFold-like networks, the pair representation stores L^2 entries for L residues, and the triangle updates take time proportional to L^3, so long chains and large complexes are limited by memory, which is one reason why large assemblies are built from predicted subcomplexes (Section 11.4). Direct-coupling models have about $L^2 \times 21^2$ coupling parameters, which must be estimated from a finite alignment. Structure search avoids expensive all-atom comparisons by encoding structures as strings, as Foldseek does (Section 9.6), at some cost in sensitivity.

9.11.5 Optimisation and energy functions

Structure prediction, docking and design are all **search problems** on an **energy landscape**: an energy or score function $E(\mathbf{x})$ is defined over conformations, ligand poses or sequences, and the task is to find its lowest minimum. Two separate things must work. The energy function must give the correct answer the lowest energy, and the search must find it. The two failures can be told apart: if the experimental structure scores better than the best model, sampling failed; if it scores worse, the energy function is wrong. In both cases the landscape is rugged and of very high dimension, with many local minima (Section 9.5).

Local optimisation, or energy minimisation, moves downhill from a starting point with gradient descent or related methods (conjugate gradients, L-BFGS) until it reaches the nearest local minimum. It removes clashes and strained geometry, for example in the final relaxation of AlphaFold2 models, but it cannot cross barriers. **Global optimisation** accepts some uphill moves. Monte Carlo with the Metropolis criterion and simulated annealing were described above (Kirkpatrick et al. 1983). **Replica exchange** runs copies of the system at several temperatures and swaps them with a Metropolis-like rule, so that conformations trapped at low temperature can escape at high temperature (Sugita and Okamoto 1999). **Genetic algorithms** evolve a population of solutions by mutation, recombination and selection, and are used in some docking programs. Discrete problems, such as choosing side-chain rotamers, have their own exact or pruning methods (see *Loops and side chains* in Section 9.8). The same ideas appear elsewhere in the book: heuristic tree search with rearrangements in phylogenetics is hill climbing in tree space, with the same risk of local optima (Section 4.3); docking searches poses (below and Section 11.4); design searches sequence space (Section 9.12); and training a neural network is gradient descent on a loss function with millions of parameters (4). In CASP13, the first AlphaFold folded proteins by gradient descent on a predicted distance potential (Senior et al. 2020).

There are two kinds of energy function. **Physics-based force fields** (see *Molecular dynamics* above) describe interactions between atoms and are transferable, but treating solvent and entropy accurately is expensive. **Knowledge-based** or **statistical potentials** are derived from known structures. The frequency of a feature, such as the distance r between two residue types, is compared with its frequency in a *reference state* that lacks specific interactions, and converted to an energy with the **inverse Boltzmann relation**:

$$\Delta E(r) = -k_B T \ln \frac{P_{\text{obs}}(r)}{P_{\text{ref}}(r)}.$$

Features that are observed more often than expected get favourable energies. Miyazawa and Jernigan derived residue–residue contact energies in this way (Miyazawa and Jernigan 1985), and Sippl derived distance-dependent **potentials of mean force** (Sippl 1990), which were used for threading (Section 9.8) (Jones et al. 1992) and for model quality assessment.

The derivation rests on assumptions that are not strictly true: that the structures in the PDB are a Boltzmann ensemble at one temperature, when they are many different proteins; that pair terms are independent and additive; and that the reference state is correct, although the result depends strongly on how it is chosen. Preferences that arise from chain connectivity or composition are also absorbed into the "energies" (Thomas and Dill 1996). Statistical potentials nevertheless work, because they summarise effective free energies, including solvation averaged over many environments, and because their task is usually to *discriminate* native-like from wrong structures, not to compute absolute stabilities. A deep network trained on the PDB can be seen as an extremely detailed learned potential of this kind.

Practical scoring functions mix both. The **Rosetta energy function** (REF2015) combines physical terms (Lennard-Jones packing, electrostatics, implicit solvation and orientation-dependent hydrogen bonds) with statistical terms (backbone torsion and rotamer probabilities, and reference energies for

each amino acid used in design), and its weights are fitted so that native structures and sequences score best and measured stabilities are reproduced (Alford et al. 2017). Its "Rosetta energy units" are roughly calibrated against kcal/mol but are not free energies.

9.11.6 Docking and scoring functions

Molecular docking predicts how a small molecule binds to a protein: its position and orientation (the *pose*) in a binding site, and, less reliably, how strongly it binds. Docking has two parts: a **search** over poses and ligand conformations, and a **scoring function** that ranks them.

- DOCK, from Irwin Kuntz's group, was the first docking program. It matched ligand atoms to spheres filling the binding site (Kuntz et al. 1982), and later added force-field-based scoring (Meng et al. 1992).
- Scoring functions are of three kinds: **force-field** based (van der Waals and electrostatic terms), **empirical** (weighted terms for hydrogen bonds, hydrophobic contacts and lost rotational entropy, fitted to measured affinities (Böhm 1994)), and **knowledge-based** (statistical potentials derived from known complexes (Muegge and Martin 1999)).
- **AutoDock Vina** combined an empirical scoring function with efficient gradient-based optimisation and became one of the most widely used tools (Trott and Olson 2009).

Docking reproduces known binding poses reasonably well, but scoring functions rank compounds by affinity poorly (Warren et al. 2005), because they neglect solvation, entropy and protein flexibility. Docking scores are often reported in kcal/mol-like units, but they are not binding free energies (Section 9.3). Ranking poses of one ligand, ranking ligands against one target, and estimating absolute affinities are different problems of increasing difficulty. More rigorous free-energy calculations with MD (free-energy perturbation, FEP) are more accurate but expensive. Still, **virtual screening** of large libraries, now billions of compounds, has found new drug leads for many targets. Protein–protein docking uses similar ideas; the classical approach used fast Fourier transforms to scan rigid shape complementarity (Katchalski-Katzir et al. 1992) (Chapter 11).

Deep learning has entered docking from two directions. Co-folding methods such as AlphaFold3 and its successors (Section 9.8) predict the protein and ligand together, and often outperform classical docking, especially when no experimental structure of the protein exists. Benchmarks such as PoseBusters showed, however, that some deep-learning methods produce physically impossible poses, with clashes and distorted geometry, and that their accuracy drops for ligands and pockets unlike those in the training data (Buttenschoen et al. 2024). Generalisation to new chemistry is the main challenge for the drug-design engines described above.

See also: further explanation in Section D.9.10; related Wikipedia articles in Section E.9.10.

Sources

Parts of this section are adapted from the Wikipedia articles Molecular dynamics, Force field (chemistry), Normal mode, Docking (molecular) and Scoring functions for docking (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Structure: Evidence, Models, and Dynamics" (CC BY 4.0), used with permission.

9.12 Protein design

Protein design reverses structure prediction: instead of finding the structure of a given sequence, it finds sequences that fold into a desired structure or carry out a desired function. It tests our understanding of proteins in the most direct way, since a design either works in the laboratory or not, and it has produced new enzymes, binders, vaccines and materials. Like prediction, it has been transformed by deep learning (Kortemme 2024).

9.12.1 Physics-based design

The first fully automated design of a protein sequence was reported by Dahiyat and Mayo in 1997. They designed a zinc-finger fold that folded without zinc, by searching over all sequences and side-chain rotamers for the lowest energy with a fixed backbone (Dahiyat and Mayo 1997). The search is combinatorial, and was solved with dead-end elimination and other optimisation methods.

David Baker's group extended this with the **Rosetta** energy function and sampling methods developed for prediction (Section 9.8). In 2003 they designed **Top7**, a protein with a fold not seen in nature, whose crystal structure matched the design almost exactly (Kuhlman et al. 2003). Over the next decade, rules for designing ideal proteins from scratch were worked out (Koga et al. 2012), and Rosetta was used to design enzymes for reactions with no natural catalyst (Jiang et al. 2008), symmetric assemblies, nanocages and vaccines (Huang et al. 2016). Most designs still failed, however, and success often required testing many designs and laboratory evolution.

9.12.2 Deep-learning design

Since 2021, design has been reorganised around deep networks, typically in two steps: generate a **backbone** structure, then design a **sequence** for it, and check the result by predicting its structure.

A central warning applies to all the methods below. A confident structure prediction of a design is an *in silico* filter: it shows that the design is consistent with what a predictor has learned, not that the sequence folds, is soluble, expresses, or binds its target specifically. Designs are often optimised against the same kind of network that is used to check them, which makes high confidence easier to obtain than real success. Only experiments validate a design.

- **Inverse folding. ProteinMPNN** is a graph neural network that takes a backbone and predicts a sequence that will fold into it (Dauparas et al. 2022). It replaced Rosetta's sequence design, is much faster, and increased the experimental success rate dramatically.
- **Hallucination.** A structure-prediction network can be run "in reverse": starting from a random sequence, the sequence is changed by Monte Carlo or gradient descent to maximise the network's confidence that it folds into a well-defined structure. Hallucination produced new folds that were confirmed experimentally (Anishchenko et al. 2021), and could build scaffolds around functional sites (Wang et al. 2022). **BindCraft** uses AlphaFold2 in this way to design protein binders against a target in a single step, with experimental success rates often from 10% to 100% (Pacesa et al. 2025).
- **Diffusion models.** Generative diffusion models, trained to remove noise from protein structures, generate new backbones from noise. **RFdiffusion**, fine-tuned from RoseTTAFold, can be conditioned on symmetry, a binding target or a functional motif, and has produced high-affinity binders, symmetric assemblies and enzymes (Watson et al. 2023). **Chroma** is a diffusion model whose generation can be steered by "programmable" conditions such as shape, symmetry or natural language (Ingraham et al. 2023). RFdiffusion All-Atom extends design to proteins that bind small molecules (Krishna et al. 2024).

- **Combining prediction and generation. Protein Hunter** exploits the tendency of diffusion-based structure predictors such as AlphaFold3 and Boltz to "hallucinate" folded structures even for poor sequences. It alternates structure prediction and sequence redesign, starting from an uninformative sequence, to converge on a sequence and structure that are consistent with each other (a preprint) (Cho et al. 2025).
- **Generative language models** such as ESM3 generate sequences directly, conditioned on structure and function (Section 9.9).

In all these pipelines, structure prediction serves as the *in silico* filter described above: designs are kept only if AlphaFold or a similar method predicts, with high confidence, that they fold into the intended structure or complex. This filter enriches for successful designs, but the most promising designs must always be tested in the laboratory. David Baker received half of the 2024 Nobel Prize in Chemistry for computational protein design.

9.12.3 Negative design and specificity

Most of these methods perform **positive design**: they make the intended structure or complex favourable. A useful protein must also avoid competing states, such as misfolded or swapped conformations, unwanted oligomers and aggregates, and, for binders, binding to off-targets such as close homologues of the target. **Negative design** makes these alternatives unfavourable, for example by also scoring a design against off-target proteins or alternative conformations and rejecting designs predicted to form them. It is essential for specific binders, for designed protein–protein interfaces that must pair only with their intended partner, and for proteins designed to switch between states. A high confidence score for the intended complex says nothing about whether the design also binds something else.

See also: further explanation in Section D.9.11; related Wikipedia articles in Section E.9.11.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Protein design and Diffusion model (CC BY-SA 4.0, retrieved 4 October 2026).

9.13 Key concepts

A checklist for revision.

- **CASP:** blind prediction; target categories; RMSD, GDT_TS, TM-score, lDDT and DockQ; CAPRI and CAMEO; key moments (Rosetta in CASP3–5, coevolution in CASP11, AlphaFold in CASP13 and CASP14, the AlphaFold3 server as a tool used by many groups in CASP16); template and training-set overlap (leakage) in benchmarks
- **Biochemical background:** the 20 amino acids and their properties; the peptide bond; free energy, the hydrophobic effect and other forces; K_d and the standard binding free energy; enzymes, transition states, Michaelis–Menten kinetics ($K_m \neq K_d$), inhibition and allostery
- **Levels of structure:** primary to quaternary; α-helix and β-sheet; Ramachandran plot; helix and sheet packing; supersecondary motifs; domains, folds and superfamilies; structural classes
- **Experimental methods:** X-ray crystallography, the phase problem and molecular replacement; resolution, R and R-free; NMR and ensembles; cryo-EM, the resolution revolution, Fourier shell correlation and conformational heterogeneity; validation; integrative structural biology (SAXS, cross-linking MS, HDX-MS, FRET/EPR; the nuclear pore complex; PDB-IHM; validation with data not used for modelling); integration with prediction

- **Folding:** Anfinsen's thermodynamic hypothesis and its exceptions; the native state as an ensemble; the hydrophobic effect and marginal stability; two-state folding; Levinthal's paradox; energy landscapes and funnels; minimal frustration; folding simulations on Anton
- **Misfolding and amyloid:** cross-β structure and steric zippers; amyloid as a generic, concentration-dependent state; aggregation-prone regions, gatekeepers, TANGO and Waltz; prions (PrP, yeast [URE3] and [PSI+]); cryo-EM structures of tau, amyloid-β and α-synuclein fibrils and polymorphs; functional amyloids
- **Structure comparison:** RMSD and its problems; TM-score; DALI, TM-align and Foldseek with the 3Di alphabet; structure usually more conserved than sequence (a tendency, not a law); the limited number of folds; clustering the AlphaFold Database
- **Structure databases:** PDB; SCOP, CATH, ECOD and TED; AlphaFold Database; ESM Metagenomic Atlas
- **Secondary structure prediction:** Chou–Fasman, GOR, PHD and PSIPRED; why alignments helped; Q3 accuracy
- **Secondary-structure assignment:** DSSP (hydrogen-bond energy, eight states) and STRIDE; reduction to three states; why assignments differ; Q3 vs. the segment overlap score (SOV)
- **Homology modelling and fold recognition:** template search, alignment, MODELLER; dependence on sequence identity; loop modelling (database, ab initio, kinematic closure); rotamers and backbone-dependent rotamer libraries; side-chain packing (dead-end elimination, SCWRL4, Rosetta packer); the Rosetta suite; deep-learning predictors model side chains implicitly; 3D profiles, threading, profile–profile methods
- **De novo prediction:** fragment assembly (Rosetta, FRAGFOLD), energy functions and sampling
- **Coevolution:** correlated mutations; direct vs. indirect correlations; DCA, pseudo-likelihood and PSICOV; contact and distance prediction with deep networks
- **AlphaFold2:** MSA and pair representations, Evoformer, triangle updates, structure module and invariant point attention, recycling, self-distillation; pLDDT, PAE and ipTM and what they do not measure; RoseTTAFold, ColabFold, OpenFold
- **Complexes and co-folding:** AlphaFold-Multimer; AlphaFold3, Pairformer and diffusion; Chai-1, Boltz-1/2 and Protenix; ESMFold2, IsoDDE and OpenDDE (2026; preprint, technical report, software); open vs. closed systems; limitations (single conformations, MSA dependence, antibodies, hallucinations, physically invalid poses, missing cofactors and modifications)
- **Intrinsically disordered proteins:** disordered regions, folding upon binding and conditional folding; disorder prediction; low pLDDT as a disorder proxy
- **Validating a predicted model:** confidence measures, stereochemistry, leakage, oligomeric state, ligands and cofactors, consistency with experiments
- **Language models:** ESM-2 and ESMFold; the ESM Metagenomic Atlas; ESM C, ESMFold2 and ESM3; ProstT5 and the 3Di alphabet; language models vs. alignments
- **Membrane proteins:** α-helical vs. β-barrel; hydrophobicity plots and their limits; positive-inside rule; TMHMM and DeepTMHMM; signal peptides; β-barrel predictors
- **Dynamics:** structure prediction vs. ensemble prediction; molecular dynamics, force fields, time scales and enhanced sampling; QM/MM; normal modes and elastic network models
- **Sampling and cost:** the Boltzmann distribution (1.4 kcal/mol ≈ tenfold population at 25 °C); Monte Carlo and the Metropolis criterion; simulated annealing; correlated frames and tests of convergence; scaling of MD (N, $N \log N$), of the pair representation (L^2 memory, L^3 triangle updates) and of DCA parameters
- **Optimisation and energy functions:** prediction, docking and design as search on an energy landscape; search vs. energy-function failures; local minimisation vs. global search (Monte Carlo, simulated annealing, replica exchange, genetic algorithms); physics-based force fields vs. knowledge-based potentials; inverse Boltzmann relation, potentials of mean force (Sippl) and contact energies (Miyazawa–Jernigan), their assumptions and why they work; the Rosetta energy function

- **Docking:** search and scoring; force-field, empirical and knowledge-based scoring functions; docking scores are not binding free energies; DOCK and AutoDock Vina; virtual screening; FEP; deep-learning docking and PoseBusters
- **Protein design:** Dahiyat and Mayo; Rosetta and Top7; ideal proteins, enzymes and nanocages; ProteinMPNN; hallucination and BindCraft; diffusion (RFdiffusion, Chroma, RFdiffusion All-Atom); Protein Hunter; ESM3; negative design and specificity; prediction as an *in silico* filter, not validation

9.14 Recommended reading

These papers are recommended for reading in full:

- Anfinsen (1973) — why the sequence determines the structure; the basis for all structure prediction.
- Marks et al. (2012) — protein structure prediction from sequence variation; coevolution explained.
- Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. (2021) — AlphaFold2; the method that largely solved single-chain structure prediction.
- Abramson et al. (2024) — AlphaFold3; co-folding of proteins with nucleic acids, ligands and ions by diffusion.
- Lin et al. (2023) — ESMFold and the ESM Metagenomic Atlas; structure from a language model.
- Watson et al. (2023) — RFdiffusion; de novo design of protein structure and function with diffusion models.

Titles, abstracts, links and the full texts are in Section A.9; a longer list of further papers is in Section B.9.

9.15 Background reading (open access)

- **Protein structure (Wikipedia)** — Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. Proteins are polymers – specifically polypeptides – formed from sequences of amino acids, which are the monomers of the polymer.
  https://en.wikipedia.org/wiki/Protein_structure
- **Protein domain (Wikipedia)** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure. Many proteins consist of several domains, and a domain may appear in a variety of different proteins.
  https://en.wikipedia.org/wiki/Protein_domain
- **Protein folding (Wikipedia)** — Protein folding is the physical process by which a protein, after synthesis by a ribosome as a linear chain of amino acids, changes from an unstable random coil into a more ordered three-dimensional structure. This structure permits the protein to become biologically functional or active.
  https://en.wikipedia.org/wiki/Protein_folding
- **Anfinsen's dogma (Wikipedia)** — Anfinsen's dogma, also known as the thermodynamic hypothesis, is a postulate in molecular biology. It states that, at least for a small globular protein in its standard physiological environment, the native structure is determined only by the protein's amino acid sequence.
  https://en.wikipedia.org/wiki/Anfinsen%27s_dogma

- **Levinthal's paradox (Wikipedia)** — Levinthal's paradox is a thought experiment in the field of computational protein structure prediction that concerns the mechanism through which peptides reach a stable native configuration, given the large amount of folding conformations.
  https://en.wikipedia.org/wiki/Levinthal%27s_paradox
- **X-ray crystallography (Wikipedia)** — X-ray crystallography is the experimental science of determining the atomic and molecular structure of a crystal, in which the structure causes a beam of incident X-rays to diffract in specific directions.
  https://en.wikipedia.org/wiki/X-ray_crystallography
- **Cryogenic electron microscopy (Wikipedia)** — Cryogenic electron microscopy (cryo-EM) is a transmission electron microscopy technique applied to samples cooled to cryogenic temperatures. Developed in the 1970s, advances in detector technology and software allow biomolecular structures to be imaged at near-atomic resolution.
  https://en.wikipedia.org/wiki/Cryogenic_electron_microscopy
- **Protein structure prediction (Wikipedia)** — Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its secondary and tertiary structure from primary structure. Structure prediction is different from the inverse problem of protein design.
  https://en.wikipedia.org/wiki/Protein_structure_prediction
- **Homology modeling (Wikipedia)** — Homology modeling, also known as comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein.
  https://en.wikipedia.org/wiki/Homology_modeling
- **Threading (Wikipedia)** — In molecular biology, protein threading, also known as fold recognition, is a method of protein modeling which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure.
  https://en.wikipedia.org/wiki/Threading_(protein_sequence)
- **AlphaFold (Wikipedia)** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.
  https://en.wikipedia.org/wiki/AlphaFold
- **CASP (Wikipedia)** — Critical Assessment of Structure Prediction (CASP), sometimes called Critical Assessment of Protein Structure Prediction, is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994.
  https://en.wikipedia.org/wiki/CASP
- **Membrane topology (Wikipedia)** — Topology of a transmembrane protein refers to locations of N- and C-termini of membrane-spanning polypeptide chain with respect to the inner or outer sides of the biological membrane occupied by the protein.
  https://en.wikipedia.org/wiki/Membrane_topology
- **Molecular dynamics (Wikipedia)** — Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system.
  https://en.wikipedia.org/wiki/Molecular_dynamics
- **Docking (molecular) (Wikipedia)** — In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when a ligand and a target are bound to each other to form a stable complex.
  https://en.wikipedia.org/wiki/Docking_(molecular)
- **Protein design (Wikipedia)** — Protein design is the rational design of new protein molecules to design novel activity, behavior, or purpose, and to advance basic understanding of protein function. Proteins can be designed from scratch or by making calculated variants of a known protein structure and its sequence.

https://en.wikipedia.org/wiki/Protein_design

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

9.16 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

9.17 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Protein Structure: Evidence, Models, and Dynamics" and "Structural AI and Generative Biology" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

9.17.1 Explain structure and folding from first principles

What determines the three-dimensional structure of a protein?

Extensions, solution outline and common errors

Extensions:

- What do the ϕ and ψ angles represent?
- Is the hydrophobic effect responsible for a fixed fraction, such as 90%, of the folding free energy?
- Why is glycine common in turns but not a general helix stabiliser?
- Does Anfinsen's hypothesis mean that every protein folds spontaneously in a cell?

Solution outline:

1. Begin with the planarity of the peptide bond, the backbone torsions, steric restrictions (the Ramachandran plot), side-chain chemistry and the solvent.
2. Explain secondary structures as recurring backbone hydrogen-bond patterns, and folds and domains as tertiary organisation.
3. Write $\Delta G = G_N - G_U = -RT \ln K$, with $K = [N]/[U]$, and separate equilibrium stability from kinetic barriers.
4. Describe a free-energy landscape with many microstates rather than a single deterministic path (funnels, Levinthal's paradox).
5. State that the hydrophobic effect, hydrogen bonds, packing, electrostatics, entropy, cofactors, membranes and partners all contribute in a context-dependent way, and reject universal percentage claims; proteins are only marginally stable.
6. Correct the glycine claim: lacking a side chain, it can adopt unusual torsions and fits turns, but its flexibility (conformational entropy) makes it a poor helix former.
7. Qualify Anfinsen's hypothesis with chaperones, kinetic traps, oligomerisation, modifications and conditional or metamorphic states.

Common errors:

- Calling the folded state “low energy” without specifying free energy, solvent and conditions.
- Treating every loop as disordered.
- Saying that one sequence always has exactly one structure.

Connections: Section 9.3 and Section 9.5; evolutionary constraints on stability in Section 3.5; ensembles in Section 9.11.

9.17.2 Choose an experimental structure method

Compare X-ray crystallography, NMR and cryo-EM for determining a protein structure.

Extensions, solution outline and common errors

Extensions:

- What is the crystallographic phase problem?
- What does an NMR ensemble represent?
- How is cryo-EM resolution estimated?
- Is a cryo-EM structure necessarily more native than a crystal structure?

Solution outline:

1. Define the observations: diffraction intensities, magnetic-resonance restraints, or many low-dose particle images.
2. Explain how each becomes a model: phasing (molecular replacement, anomalous scattering), map fitting and refinement; calculation of conformers that satisfy the restraints; particle alignment, classification, reconstruction and map fitting.
3. Compare sample requirements and strengths: crystals and detailed density; solution restraints and dynamics for smaller proteins; large non-crystalline assemblies and several conformational classes.
4. Name method-specific failures: crystal contacts and model bias; size limits and incomplete restraints; preferred orientation, radiation damage and uneven local resolution.
5. Validate with held-out data (R_{free}), restraint violations, or independent half-maps (Fourier shell correlation) plus map–model fit, and check stereochemistry for all.
6. Choose by the biological question, and combine orthogonal data (integrative modelling) when one method cannot distinguish the alternatives.

Common errors:

- Equating nominal resolution with uniform accuracy of every atom.
- Reading the spread of an NMR ensemble as measured thermodynamic populations.
- Claiming that vitrification guarantees an unperturbed native ensemble.

Connections: Section 9.4; the PDB and validation reports in Section 2.5; cross-linking MS in Section 8.5.

9.17.3 Compare classical prediction strategies

How do comparative (homology) modelling, threading, de novo (fragment- or physics-based) prediction and coevolution-based prediction differ?

Extensions, solution outline and common errors

Extensions:

- Where do comparative models usually fail?
- What makes de novo prediction computationally difficult?
- Why are raw correlations between alignment columns not direct contacts?
- What happens when the MSA is shallow?

Solution outline:

1. Comparative modelling transfers spatial restraints from a homologous template through an alignment (MODELLER); errors in the choice of template or state and in the alignment dominate.
2. Threading and fold recognition test how well a sequence fits known folds when sequence similarity is weak.
3. Fragment assembly and physics-based methods sample conformations under approximate energy or scoring functions; both the search and the ranking can fail.
4. Coevolution methods fit a global (Potts) model to a deep, diverse MSA to separate direct from indirect correlations (DCA, PSICOV) and derive contact restraints.
5. Compare information sources, computational cost, interpretability and failure modes rather than claiming one universal hierarchy.
6. Validate with blind targets (CASP), local geometry, domain-wise scores and independent experiments.

Common errors:

- Defining homology modelling by a fixed percentage identity threshold.
- Saying that de novo prediction enumerates every conformation with exact physics.
- Treating a strong coupling as proof of a contact within one chain (it may be between chains, or reflect alternative conformations).

Connections: Section 9.8; profiles and alignments in Section 1.6; phylogenetic confounding of coevolution in Chapter 4.

9.17.4 Explain AlphaFold2 and its confidence

At a high level, how does AlphaFold2 predict structure, and how should its output be evaluated?

Extensions, solution outline and common errors

Extensions:

- What information do the Evoformer and the structure module exchange?
- Contrast pLDDT, PAE and pTM.
- Can a high pLDDT prove a catalytic mechanism?
- How would you assess a predicted two-domain protein?

Solution outline:

1. Describe the inputs (sequence, MSA and optional templates); the Evoformer jointly updates the MSA and residue-pair representations, with triangle updates enforcing geometric consistency.
2. Explain how the structure module generates coordinates with invariant point attention, and the role of recycling.
3. Interpret pLDDT as local confidence, PAE as confidence in the relative placement of residues and domains, and pTM as a global score.
4. For a multidomain protein, inspect the pLDDT of each domain and the off-diagonal blocks of the PAE matrix: high local confidence can coexist with uncertain packing of the domains.
5. Check the MSA and templates, possible training-set leakage, alternative seeds, stereochemistry, disorder and the biological state (ligands, cofactors, oligomers).
6. Treat the model as a hypothesis, and validate the intended fold, active site or state independently.

Common errors:

- Calling pLDDT the probability that the structure is correct.
- Reading a low-confidence ribbon as an atomic model of a disordered ensemble.
- Inferring substrate specificity from a confident global fold.

Connections: Section 9.8; language-model alternatives in Section 9.9; function transfer in Section 10.4; learned confidence and leakage in 4.

9.17.5 Evaluate AlphaFold-Multimer and AlphaFold3 complexes

AlphaFold predicts a high-confidence protein–ligand or protein–protein complex. What does that result establish?

Extensions, solution outline and common errors

Extensions:

- What changed from AlphaFold2 to AlphaFold-Multimer and AlphaFold3?
- Which confidence measures matter at an interface?
- Does AlphaFold3 predict affinity?
- Which errors specific to co-folding models should you look for?

Solution outline:

1. AlphaFold-Multimer added training on complexes and interface confidence (ipTM); AlphaFold3 represents proteins, nucleic acids, ligands and ions as tokens and uses a Pairformer with a diffusion module that generates atomic coordinates.
2. Confirm the construct, stoichiometry, ligands, modifications, ions and biological state before prediction.
3. Inspect inter-chain PAE, ipTM, interface pLDDT, convergence across seeds and alternative partners or states.
4. Check stereochemistry, chirality, clashes, protonation, metals and the plausibility of the ligand pose (PoseBusters).
5. State explicitly that giving two sequences as input can force a plausible interface, and that confidence is not K_d, kinetics, occupancy or evidence of interaction in the cell.

6. Test a discriminating prediction about the interface with binding measurements, mutations with folding controls, cross-links, cellular perturbation or an experimental structure.

Common errors:

- Calling two proteins interactors because a server returned a complex.
- Treating the top-ranked ligand pose as a measured bound state.
- Ignoring static-state bias, hallucinated order, chirality errors and clashes.

Connections: Section 9.8; interaction assays and complex prediction in Section 11.2 and Section 11.4; out-of-distribution prediction in 4.

9.17.6 Select and interpret structural accuracy metrics

Compare RMSD, GDT_TS, TM-score, lDDT and DockQ.

Extensions, solution outline and common errors

Extensions:

- Why can two accurately predicted domains give a poor whole-chain RMSD?
- Which measures need a global superposition?
- What does lDDT compare?
- How would you score a predicted ligand-binding site?

Solution outline:

1. RMSD averages squared distances between matched atoms after a chosen superposition, and is sensitive to outliers and domain motion.
2. GDT_TS reports the fractions of residues within several distance cutoffs (1, 2, 4 and 8 Å), and is more robust to a single badly placed region.
3. TM-score normalises by length and emphasises the global topology.
4. lDDT compares local inter-atomic distances without a global superposition, and tolerates relative domain motion.
5. DockQ combines the fraction of native contacts and interface and ligand RMSDs for complexes.
6. Match the measure to the claim (global topology, domains, interface, ligand pose or active-site geometry), and always report coverage and residue correspondence.

Common errors:

- Quoting universal "good" RMSD thresholds independent of size and alignment.
- Comparing scores computed over different residue subsets without saying so.
- Assuming that a correct global topology ensures accurate side chains or ligands.

Connections: Section 9.2 and Section 9.6; interface quality in Section 11.4; choosing evaluation measures in 4.

9.17.7 Use docking and molecular dynamics responsibly

How should docking and molecular dynamics be combined to study ligand binding?

Extensions, solution outline and common errors

Extensions:

- What are the search and scoring problems in docking?
- Why might docking into an apo structure fail?
- Can ordinary MD show the formation of a covalent bond?
- How do you show that a set of trajectories has converged?

Solution outline:

1. Define the receptor state, the protonation and tautomer of the ligand, cofactors, waters, membrane, and whether binding is expected to be covalent.
2. Dock by sampling translations, rotations, ligand torsions and selected receptor flexibility, then rank poses with an approximate scoring function.
3. Validate docking by redocking and cross-docking known ligands, decoys, alternative receptor states and experimental data.
4. Parameterise and equilibrate MD carefully; standard MD integrates Newton's equations under an empirical force field and does not make or break covalent bonds (QM/MM is needed for reactions).
5. Use independent simulations from different starts, block averages and comparison with experimental observables; distinguish sampling failure from force-field error.
6. Report mechanistic hypotheses, not a trajectory as proof. Affinities require free-energy methods (such as FEP) with uncertainties, not a raw docking score.

Common errors:

- Defining a ligand as something that binds irreversibly.
- Assuming that a longer simulation automatically corrects a wrong protonation or conformational state.
- Describing one spontaneous-looking event as a rate or an equilibrium population.

Connections: Section 9.11 and Section 9.3; biophysical validation of binding in Section 11.2.

9.17.8 Design a protein and validate it

Describe a modern computational protein-design workflow, and specify how its products should be validated.

Extensions, solution outline and common errors

Extensions:

- How do backbone generation and inverse folding differ?
- What do RFdiffusion and ProteinMPNN contribute?
- Why is AlphaFold confidence insufficient as validation?
- How would you detect memorisation of training structures or an unwanted oligomer?

Solution outline:

1. Specify the design objective and the negative constraints: topology, motif geometry, partner, affinity, specificity, solubility and oligomeric state.

2. Generate or scaffold candidate backbones (for example with RFdiffusion), and use inverse folding (ProteinMPNN) to propose sequences compatible with each backbone.
3. Filter with independent structure prediction, checks for alternative states and oligomers, energy and geometry, sequence diversity, and aggregation.
4. Search structure and sequence databases to assess novelty and similarity to training structures.
5. Test experimentally: expression, monodispersity, folding, stability, oligomeric state, structure, affinity and kinetics, specificity, and function.
6. Feed failures back into the design while keeping held-out assays as genuine validation.

Common errors:

- Calling a design with a high-confidence prediction experimentally validated.
- Optimising only for binding and ignoring specificity (negative design) or aggregation.
- Testing one design and generalising its success to the method.

Connections: Section 9.12; structure search for novelty in Section 9.6; binding assays in Section 11.2; generative models in Section 12.5.

9.17.9 Statistical potentials and the search for the native structure

In a set of non-redundant protein structures, leucine–leucine contacts are observed 2.5 times more often than expected from a reference state. Convert this to an energy, explain the assumptions behind the conversion, and discuss why a de novo prediction with such a potential can fail.

Extensions, solution outline and common errors

Extensions:

- Why does the result depend on the choice of reference state?
- How would you decide whether a failed prediction was caused by the energy function or by the search?
- How do local minimisation, simulated annealing and replica exchange differ?
- In what sense is a network such as AlphaFold a learned statistical potential, and in what sense is it not?

Solution outline:

1. Apply the inverse Boltzmann relation, $\Delta E = -k_B T \ln(P_{\text{obs}}/P_{\text{ref}})$; with $k_B T \approx 0.6$ kcal/mol at 300 K, $\Delta E \approx -0.6 \ln 2.5 \approx -0.55$ kcal/mol, a favourable contact.
2. State the assumptions: the structures behave like a Boltzmann ensemble at one temperature, pair terms are independent and additive, and the reference state removes everything but the specific interaction. None holds exactly: the database is many different proteins, and connectivity and composition leak into the "energies" (Thomas and Dill).
3. Explain why they still work: they are effective free energies that include solvation, and they are mainly used to discriminate native-like from wrong structures (threading, model quality assessment).
4. Separate the two failure modes: compare the energy of the experimental structure (after minimisation) with that of the best model; if the native scores better, sampling failed; if worse, the energy function failed.

5. Describe the search: local gradient-based minimisation finds only the nearest minimum; Monte Carlo, simulated annealing and replica exchange accept uphill moves to cross barriers; fragment assembly reduces the search space.
6. Relate to deep learning: networks learn far richer, context-dependent preferences from the PDB and output a structure directly, but they still reflect the biases of their training data.

Common errors:

- Treating a statistical potential as a physical free energy of binding or folding.
- Assuming that a lower score always means a more correct model, without checking for search failure.
- Forgetting that more sampling cannot fix an energy function that prefers a wrong structure.

Connections: Section 9.11 (optimisation and energy functions); fragment assembly and threading in Section 9.8; heuristic tree search in Section 4.3; gradient descent in 4.

9.17.10 Amyloid and aggregation

A patient carries a missense mutation in a secreted protein that replaces a lysine with an isoleucine. The protein is found in extracellular fibrils. Explain how the mutation could cause disease, and how you would use computational and structural methods to investigate it.

Extensions, solution outline and common errors

Extensions:

- What is the cross-β structure, and how is it detected?
- What is a prion, and how does it differ from other amyloids?
- Why would AlphaFold not show the fibril structure?
- Give an example of a functional amyloid.

Solution outline:

1. Describe amyloid: cross-β fibrils with strands perpendicular to the axis, about 4.7 Å and 10 Å spacings, steric zippers; a generic, concentration-dependent state that competes with the native fold.
2. Consider two mechanisms: the mutation destabilises the native fold, increasing the population of unfolded or partly folded chains, or it strengthens an aggregation-prone region by removing a charged gatekeeper and increasing hydrophobicity and β-propensity.
3. Predict aggregation-prone regions with TANGO or Waltz for the wild type and the mutant, and predict the stability change of the native structure (Section 10.6).
4. Check the location of the residue in an experimental or predicted native structure: buried or exposed, in a strand or a gatekeeper position.
5. Determine the fibril structure by cryo-EM, preferably from patient tissue, since in vitro fibrils can be different polymorphs; test aggregation kinetics experimentally.
6. Mention prions (PrP, yeast [URE3] and [PSI+]) as aggregates that template conversion and spread, and functional amyloids such as curli.

Common errors:

- Assuming that a predicted native structure describes the aggregated state.
- Assuming that all amyloid is pathological.
- Treating one in vitro fibril structure as the disease structure.

Connections: Section 9.5 (misfolding, amyloid and prions); cryo-EM in Section 9.4; variant effects in Section 10.6.

10 Protein function

> **From the exam description**
>
> Classifications, prediction methods, and definitions.

10.1 Overview

Predicting what a protein does is the ultimate goal of most sequence analysis. It is also a broad and messy topic, because "function" can mean many things, the data are incomplete and biased, and almost every method in this book contributes to it. This chapter cannot cover everything; it concentrates on the main ideas:

- what function is and how it is described: the Gene Ontology, evidence codes, the structure of GO annotations and EC numbers (Section 10.2);
- how function is determined experimentally: forward and reverse genetics, knockout collections, RNAi and CRISPR screens, DepMap, and why many knockouts have no visible phenotype (Section 10.3);
- how function is predicted: transfer by homology and its pitfalls, families and domains, genomic context, networks, structure, functional and modification sites, disordered regions and machine learning, and how evidence is combined (Section 10.4);
- subcellular localisation and the sorting signals that determine it, and how such predictors are evaluated (3);
- predicting the effects of missense variants (Section 10.6);
- how function prediction is evaluated, and why that is hard (Section 10.7).

The databases used for function annotation are described in Section 2.6, protein families and domains in Section 1.9, and interaction networks in Chapter 11.

10.2 Definitions of function

Predicting what a protein does is the ultimate goal of most sequence analysis, yet "function" has no single definition. It can mean the biochemical activity of a protein ("serine protease"), the process it contributes to ("blood coagulation"), the place where it acts ("extracellular space"), its phenotypic effect when mutated, or the role it plays in a pathway or complex. Function is also context dependent: the same protein can have different roles in different cells, and some proteins have several unrelated functions (*moonlighting*), while many enzymes catalyse side reactions besides their main one (*promiscuity*). Any computational definition must choose which of these aspects to describe, and with what vocabulary. Two systems dominate: the Gene Ontology and the Enzyme Commission numbers.

10.2.1 The Gene Ontology

The **Gene Ontology** (GO) (Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023), described as a resource in Section 2.6, describes gene products along three separate aspects:

- **Molecular function** (MF): the activity of the gene product itself, such as "protein kinase activity" or "DNA binding";
- **Biological process** (BP): the larger programme the activity contributes to, such as "DNA repair" or "T cell activation";
- **Cellular component** (CC): where the activity takes place, such as "nucleus" or "mitochondrial inner membrane".

These are three independent dimensions of annotation, not three definitions of the same thing. Only molecular function describes what the protein itself does biochemically; a protein can be annotated in all three aspects, and a process or location annotation says little about its activity.

GO has tens of thousands of terms (roughly 40,000 non-obsolete terms in 2026), organised as a **directed acyclic graph** in which more specific terms are linked to more general ones by relations such as *is a* and *part of*. Under the *true path rule*, an annotation to a term propagates to its ancestors along these relations (an annotation to "serine-type endopeptidase activity" also implies "peptidase activity"). This holds for *is a* and *part of*, but not for every relation: a protein annotated to "regulation of" a process is not thereby involved in the process itself. Predictions can be made at different levels of specificity: "hydrolase activity" is easy to predict and nearly useless, "serine-type endopeptidase activity" is harder and more informative.

Every annotation carries an **evidence code** that records its source: experimental codes (such as IDA, *inferred from direct assay*, and IMP, *inferred from mutant phenotype*), phylogenetic and sequence-based codes (IBA, ISS), author and curator statements, and **IEA** (*inferred from electronic annotation*), assigned automatically without review. Only a very small fraction of the sequences in UniProtKB, well below 1% in 2026, have a GO annotation supported by an experimental evidence code; the exact figure depends on the release and on which evidence codes are counted as experimental. The other proteins have only predicted annotations, if any, mostly from the methods described in Section 10.4. Experimental annotations are themselves biased. A handful of model organisms and well-studied proteins account for most of them, and a few high-throughput experiments contribute many annotations to general terms (Schnoes et al. 2013). These biases matter both for training predictors and for evaluating them (Section 10.7).

The kinds of evidence are not equivalent. A **direct assay** (IDA), such as an enzyme assay of the purified protein, shows a molecular function under the assay conditions, but not that the protein performs it in the cell. A **mutant phenotype** (IMP) shows that the gene is needed for a process, but not what the protein does in it; the effect may be indirect. **Physical interactions** (IPI) and **expression patterns** (IEP) show association, not activity. **Computational transfer** (ISS, IBA, IEA) is a prediction whose reliability depends on the method and on the evidence for the source protein. A strong claim about function therefore usually combines several kinds of evidence.

10.2.2 The structure of a GO annotation

A GO annotation is more than a link between a protein and a term. It is a statement with several parts: the gene product (sometimes a specific isoform), a **relation** such as *enables* (for molecular function), *involved in* (for process) or *located in* or *part of* (for component), the GO term, the evidence code, the reference that supports it, and the group that made the annotation. Annotations can also carry:

- a **NOT qualifier**, an explicit statement, based on evidence, that the gene product does *not* have a function or location. This is different from the absence of an annotation, and it does not propagate to ancestor terms: a protein that is NOT a kinase may still have transferase activity;
- **annotation extensions**, which add context, for example the substrate of an enzyme, the cell type in which a process occurs, or the target of a regulator.

GO follows an **open-world assumption**: a missing annotation means that nothing has been recorded, not that the protein lacks the function. This is why predictions that are absent from the database cannot simply be counted as wrong (Section 10.7). Annotations change with every release, as terms are added, merged or made obsolete, so analyses should record the ontology and annotation versions. Standard annotations describe each gene product separately. **GO-CAM** models link several annotations into a causal model of a pathway, in which one activity regulates or provides input to another, in a specified location and process (Thomas et al. 2019).

10.2.3 EC numbers

Enzymes are classified by the **Enzyme Commission** (EC) of the International Union of Biochemistry, established in 1955, whose first list appeared in 1961. An EC number has four levels, for example EC 3.4.21.4 for trypsin:

1. the **class** of reaction: 1 oxidoreductases, 2 transferases, 3 hydrolases, 4 lyases, 5 isomerases, 6 ligases and 7 translocases;
2. the **subclass**, for example 3.4, hydrolases acting on peptide bonds (peptidases);
3. the **sub-subclass**, for example 3.4.21, serine endopeptidases;
4. the **serial number** of the specific reaction, for example 3.4.21.4, trypsin.

In 2026 there are more than 6,000 active EC numbers. Importantly, the EC system classifies **reactions**, not proteins: unrelated enzymes that catalyse the same reaction get the same number (*analogous* enzymes, Section 3.4), and homologous enzymes in the same family can have different numbers. Detailed data on each enzyme, such as kinetic parameters, substrates, inhibitors and organism, are collected in BRENDA (Chang et al. 2020), and the reactions themselves are described with chemical structures in Rhea (Bansal et al. 2021), which UniProt now uses for enzyme annotation.

10.2.4 Other descriptions of function

Function can also be described by membership in **families** and **pathways**: Pfam and InterPro families (Section 1.9), KEGG Orthology groups and Reactome pathways (Section 2.6), or by **interactions** with other proteins (Chapter 11). For human genes, **phenotypes** and diseases associated with mutations (for example in OMIM) describe function at the level of the organism. Each description is useful for different purposes, and a complete picture of a protein's function combines several of them.

See also: further explanation in Section D.10.1; related Wikipedia articles in Section E.10.1.

i Sources

Parts of this section are adapted from the Wikipedia articles Protein function prediction, Gene Ontology, Enzyme Commission number and Moonlighting protein (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Function and Annotation" (CC BY 4.0), used with permission.

10.3 Experimental determination of function

Every prediction method in this chapter is ultimately trained on, or transfers from, functions that were determined by experiment. The most general experimental approach is genetic: change or remove a gene and observe what happens to the organism or cell. Large-scale genetic screens are also a major source of data for bioinformatics, and analysing them is a bioinformatics problem in its own right.

10.3.1 Forward and reverse genetics

Forward genetics starts from a phenotype and looks for the genes responsible. Organisms are mutagenised at random, with chemicals, radiation or transposons, and individuals with an interesting phenotype are selected. The classical screens for mutations that disrupt the body plan of the *Drosophila* embryo identified many of the genes that control development (Nüsslein-Volhard and Wieschaus 1980). The difficult step used to be finding the causal mutation, which required genetic mapping by crosses. Today the mutant genomes are sequenced, and the causal variant is found among the many induced mutations by combining mapping with sequencing, for example by looking for the region in which all mutant offspring share the same variants. Forward screens are unbiased: they can find genes that nobody suspected.

Reverse genetics goes the other way: it starts from a gene, usually identified by sequencing, and disrupts it to see what happens. Since genome sequencing made all genes known, reverse genetics has dominated, and the question "what does this gene do?" is often answered by systematically removing every gene.

10.3.2 Genome-wide knockout collections

The first systematic collection was the **yeast deletion collection**, in which each of the roughly 6000 open reading frames of *Saccharomyces cerevisiae* was replaced by a marker cassette (Winzeler et al. 1999; Giaever et al. 2002). Each deletion carries two unique 20-nucleotide **barcodes**, so that thousands of strains can be grown together in one culture and the abundance of each strain measured by counting its barcode, first with microarrays and later by sequencing. About 1000 genes (around 19%) were essential for growth in rich medium. Pooled growth of barcoded mutants became the model for later CRISPR screens, and the same collection was used to map genetic interactions (Section 11.2).

In mouse, the International Mouse Phenotyping Consortium (IMPC) aims to knock out every protein-coding gene and phenotype the mice in a standardised pipeline. Of the first 1751 genes, about a quarter were lethal when homozygous and more were subviable, and many genes without previous annotation produced phenotypes (Dickinson et al. 2016); overall, roughly 30% of mouse gene knockouts are lethal before or around birth, which is why embryonic phenotyping became part of the pipeline (Adams et al. 2013). Knockout phenotypes are a major source of experimental GO annotations for biological process (Section 10.2).

10.3.3 RNAi and CRISPR

In other organisms, and in human cells, genes are perturbed by interfering with their expression rather than by homologous recombination:

- **RNA interference** (RNAi) uses double-stranded RNA, or short interfering RNAs (siRNAs) and short hairpin RNAs (shRNAs), to degrade the matching mRNA (Fire et al. 1998). It produces a *knockdown*: expression is reduced, usually incompletely. Its main problem is **off-target effects**: an siRNA can act like a microRNA and repress hundreds of transcripts that match its seed region (Jackson et al. 2003). Many hits of early RNAi screens were artefacts of this kind.
- **CRISPR–Cas9** knockouts use a guide RNA (sgRNA) to direct Cas9 to a target site, where it cuts both DNA strands. Error-prone repair introduces small insertions or deletions that usually shift the reading frame and produce a *knockout*. Off-target cutting at similar sites occurs, but is rarer than RNAi off-target effects.
- **CRISPR interference** (CRISPRi) and **activation** (CRISPRa) use a catalytically dead Cas9 fused to a repressor or an activator domain. It binds near a promoter without cutting, and turns expression down or up (Gilbert et al. 2014). CRISPRa allows gain-of-function screens, which are impossible with knockouts.

Knockdown and knockout phenotypes can disagree for several reasons. A knockdown leaves some protein, which may be enough. A knockout may be compensated: in zebrafish, mutants of some genes were normal while morpholino knockdowns of the same genes were not, because the mutation induced the expression of related genes (Rossi et al. 2015). Off-target effects can create phenotypes in either method, and an in-frame indel may leave a partly functional protein. Agreement between independent reagents, such as several guides or siRNAs against the same gene, and rescue by re-expressing the gene, are therefore the standard of evidence.

10.3.4 Pooled CRISPR screens

In a **pooled screen**, a library of lentiviruses carrying several guides per gene for every gene in the genome is introduced into a population of cells at a low dose, so that each cell receives about one guide. The cells are grown, or selected for a phenotype such as drug resistance, and the guides are counted by sequencing them from genomic DNA before and after (Shalem et al. 2014; T. Wang et al. 2014). The integrated guide acts as a barcode, as in the yeast deletion collection. Guides against genes needed for growth drop out of the population; guides against genes whose loss confers resistance become enriched.

The analysis resembles that of RNA-seq counts (Section 7.4):

1. Guide counts are normalised for sequencing depth, and a log fold change is computed for each guide between the end and the start of the experiment.
2. The guide-level changes are combined into a **gene-level score**. Since some guides do not cut efficiently, robust methods are used. MAGeCK models the counts with a negative binomial distribution and ranks genes with a modified robust rank aggregation, which asks whether a gene's guides are ranked higher than expected by chance (Li et al. 2014).
3. Genes are called significant with a false discovery rate, and the quality of the screen is assessed by how well it separates reference sets of known essential and non-essential genes (Hart et al. 2015).

An important artefact concerns **copy number**. Cas9 cutting at many sites in an amplified region causes DNA damage and kills cells regardless of which gene is targeted, so genes in amplified regions appear essential (Aguirre et al. 2016). Methods such as CERES and Chronos model this effect, and Chronos also models the dynamics of cell growth and the delay between cutting and loss of the protein (Meyers et al. 2017; Dempster et al. 2021).

Screens across many cell lines distinguish **core essential genes**, needed in every cell line, from *context-specific dependencies*, needed only in some (Hart et al. 2015). The **Cancer Dependency Map** (DepMap) has screened around a thousand cancer cell lines, first with RNAi and then with CRISPR, and relates the dependencies to the mutations, expression and copy number of each line (Tsherniak et al. 2017).

A gene on which only lines with a particular mutation depend is a candidate drug target (Section 7.7). Genes with similar dependency profiles across cell lines tend to work together, so DepMap is also used to predict function by co-essentiality, an idea related to the genetic interaction profiles of Section 11.2.

10.3.5 Arrayed, imaging and single-cell screens

Pooled screens can only measure phenotypes that change the number of cells carrying a guide, through growth or sorting. In an **arrayed screen**, each well of a plate receives one perturbation, so any phenotype can be measured, including cell shape and the localisation of proteins under a microscope. Image-based profiling turns such images into feature vectors that can be compared between perturbations (Section 13.4). **Perturb-seq** instead reads out both the guide and the whole transcriptome of each cell in a pooled screen, so that the transcriptional effect of every perturbation is measured (Section 11.2, Section 7.6). Such high-content screens give a rich phenotype for every gene, which can be used to group genes into pathways and complexes (Bock et al. 2022).

10.3.6 Why many knockouts show no phenotype

A large fraction of genes show no visible phenotype when they are knocked out, even in a carefully phenotyped organism. In yeast, most deletions grow normally in rich medium, although nearly all of them showed a growth defect under at least one of more than a thousand chemical and environmental conditions (Hillenmeyer et al. 2008). Several explanations apply:

- **Redundancy.** A paralogue or an alternative pathway can take over. Such genes often show synthetic lethality with their partner (Section 11.2).
- **Condition dependence.** The gene may only be needed in a condition that was not tested, such as a particular nutrient, stress or developmental stage.
- **Small effects.** A fitness defect of a few per cent is invisible in a single experiment, but matters over evolutionary time.
- **Pleiotropy.** Conversely, a gene with many roles can give a complex phenotype, such as lethality early in development, that hides its later functions.

The absence of a phenotype is therefore not evidence that a gene has no function. The phenotype that is observed is also not the molecular function: a growth defect says that a gene matters, not what its protein does. Experimental annotations from screens therefore mostly concern biological processes, and are best combined with the prediction methods of Section 10.4.

Systematic mutation can also be applied within a single protein: **deep mutational scanning** measures the effect of every possible substitution, and is described with the prediction of variant effects in Section 10.6.

See also: further explanation in Section D.10.6; related Wikipedia articles in Section E.10.6.

> **Sources**
>
> Written for this book from the primary literature cited in the text.

10.4 Function prediction

Since only a tiny fraction of proteins has been studied experimentally, the function of almost all proteins is predicted. The methods use different kinds of information: similarity to characterised proteins, membership in families, the genomic and network context of a gene, and structure. They are increasingly combined with machine learning.

10.4.1 Annotation transfer by homology

The most common approach is to transfer the annotation of the most similar characterised protein (Whisstock and Lesk 2003). It rests on the assumption that homologous proteins have similar functions, which holds approximately but with important exceptions:

- **Sequence identity thresholds.** Function is more conserved between closely related proteins. Studies of enzymes found that the first three digits of the EC number were usually conserved above about 40% sequence identity, but that the full EC number could differ even above 50% identity (Devos and Valencia 2000; Rost 2002; Tian and Skolnick 2003). Below about 30%, in the twilight zone (1), specific functions are often not shared, and usually only more general functions are conserved. These figures are empirical tendencies in particular data sets, not decision rules. There is no universal identity threshold for safe transfer: it also depends on the length and coverage of the alignment, the protein family, whether active-site residues are conserved, the domain architecture, and how specific the transferred function is.
- **Domain architecture.** Multidomain proteins are common, especially in eukaryotes. A BLAST hit that covers one domain says nothing about the others, so transferring the annotation of a whole protein from a partial match is a common source of errors. Annotation should consider whether the domain architectures match (Section 1.9).
- **Orthologues vs. paralogues.** Orthologues, separated by speciation, are expected to keep the same function more often than paralogues, which may have diverged after duplication (Section 3.4). This *ortholog conjecture* is widely used, for example in eggNOG-mapper, which transfers annotations from precomputed orthologous groups (Cantalapiedra et al. 2021), although the evidence for it is weaker than often assumed (Nehrt et al. 2011; Altenhoff et al. 2012). Orthology is useful evidence, not a guarantee: orthologues can change function, and many paralogues keep the same function. It should not be read as a rule that transfer from orthologues is safe and transfer from paralogues is not.
- **Error propagation.** Annotations transferred from proteins whose own annotations were predicted can spread errors through databases (Gilks et al. 2002). Analyses of enzyme superfamilies found high rates of misannotation in automatically annotated databases (Schnoes et al. 2009), which is why evidence codes and the provenance of annotations matter (Section 10.2).

10.4.2 Family- and domain-based annotation

Rather than transferring function from a single best hit, family-based methods assign proteins to curated families with profile HMMs and other signatures, and attach annotations to the family or domain (Section 1.9). **InterPro** integrates Pfam and the other member databases, and maps many entries to GO terms (InterPro2GO) (Blum et al. 2024). More specific resources define families of proteins with the same full-length function and curated annotation rules, used by UniProt to annotate unreviewed entries automatically. The CATH functional families (FunFams) divide structural superfamilies into functionally coherent groups. Deep learning models can also classify sequences into Pfam families directly (Bileschi et al. 2022).

10.4.3 Genomic context

Comparative genomics provides information that does not depend on similarity to a characterised protein. Three kinds of genomic context suggest that two proteins are **functionally linked**, typically in the same pathway or complex (Marcotte, Pellegrini, Ng, et al. 1999; Enright and Ouzounis 2001; Dandekar 1998; Overbeek et al. 1999; Pellegrini et al. 1999):

- **Gene fusion** (the *Rosetta stone* method): two proteins that are separate in one organism are fused into a single chain in another, which suggests that they interact (Marcotte, Pellegrini, Ng, et al. 1999; Enright and Ouzounis 2001).
- **Conserved gene neighbourhood**: genes that are adjacent in many distantly related prokaryotic genomes, often in operons, tend to work together (Dandekar 1998; Overbeek et al. 1999).
- **Phylogenetic profiles**: proteins that are present or absent together across many genomes are likely to be needed together (Pellegrini et al. 1999).

These methods predict links, not functions: the function of an uncharacterised protein is inferred from its known partners. They work best for prokaryotes, where gene order is conserved and many genomes are available, and they were among the first sources of evidence integrated into STRING (Chapter 11).

10.4.4 Networks: guilt by association

More generally, proteins that interact physically, are co-expressed, or share genetic interactions tend to share functions. **Network-based prediction** propagates annotations through such networks (Sharan et al. 2007; Marcotte, Pellegrini, Thompson, et al. 1999). The simplest version assigns to a protein the most common functions of its neighbours (*guilt by association*). More advanced methods consider the whole network, for example by diffusion, label propagation or graph neural networks. The networks themselves are described in Chapter 11. Predicted structures of complexes have made it possible to add structural evidence to interactome-based annotation of whole proteomes (Garzón et al. 2016).

10.4.5 Structure-based prediction

Structure is more conserved than sequence (Section 9.6), so structural similarity can reveal homology, and therefore possibly function, when sequence similarity is lost. Structure also shows features directly related to function: binding pockets, conserved surface patches and catalytic sites. Methods compare a structure with libraries of **active-site templates**, such as the catalytic triad of serine proteases, or combine several structure-based analyses (Laskowski et al. 2005). Graph neural networks predict GO terms from the contact graph of a structure together with language-model features (Gligorijević et al. 2021). With AlphaFold models of almost all proteins and fast search with Foldseek (Section 9.6), structure-based annotation can now be applied to whole proteomes. Its limits are those of homology transfer: the same fold often carries different functions.

10.4.6 Predicting functional sites

Function often depends on a few residues: the catalytic residues of an enzyme, the residues that bind a ligand or metal, or a modification site. Predicting these sites is useful in itself, for example to choose residues for mutagenesis, and it makes annotation transfer more reliable, since a homologue that has lost a catalytic residue has probably lost the activity. Three kinds of information are used:

- **Sequence motifs.** Short patterns and profiles, such as those in PROSITE (Sigrist et al. 2012), describe catalytic sites, binding sites and modification sites (Section 1.9). Short patterns match by chance, so a motif hit alone is weak evidence.
- **Conservation.** Functional residues are usually among the most conserved positions in a multiple alignment of a family. Conservation also reflects structural constraints, so conserved residues on the surface, or residues conserved within a subfamily but different between subfamilies (*specificity-determining positions*), are more informative than conservation alone (Ashkenazy et al. 2016).
- **Structure.** Geometric methods detect pockets and clefts that can bind ligands (Le Guilloux et al. 2009), and catalytic sites can be recognised by matching the three-dimensional arrangement of residues to templates of known active sites, such as those collected in the Mechanism and Catalytic Site Atlas (M-CSA) (Ribeiro et al. 2017). Combining pockets with conservation is more reliable than either alone, and predicted structures make these methods applicable to almost any protein.

10.4.7 Predicting post-translational modification sites

Post-translational modifications (PTMs) are a special kind of functional site: a residue that is covalently modified, often by an enzyme that recognises a short sequence around it. Predictors exist for the most common types:

- **Phosphorylation.** Kinases phosphorylate serine, threonine and tyrosine. NetPhos used neural networks trained on known sites to predict whether a residue is phosphorylated (Blom et al. 1999), and NetPhosK and later methods such as GPS predict *which kinase* modifies it (Blom et al. 2004). Kinase-specific prediction rests on the preferred sequence around the site, for example arginine at position −3 for basophilic kinases such as PKA, or proline at +1 for proline-directed kinases such as CDKs and MAP kinases. Substrate preferences measured with peptide libraries for almost all human serine/threonine kinases now give a systematic basis for such predictions (Johnson et al. 2023). Deep-learning methods predict many PTM types from sequence (D. Wang et al. 2020).
- **Glycosylation.** N-linked glycans are attached to asparagine in the *sequon* N-X-S/T (X not proline), but only in proteins that enter the secretory pathway, and only on the side of the membrane facing the lumen. NetNGlyc predicts which sequons are used. O-linked glycosylation on serine and threonine has no simple motif, and NetOGlyc was trained on sites mapped by mass spectrometry in engineered cell lines (Steentoft et al. 2013).
- **Lipidation.** N-myristoylation of an N-terminal glycine, S-palmitoylation of cysteines, prenylation of a C-terminal CaaX box and the GPI anchor

  (3) attach proteins to membranes, and have their own predictors (Xie et al. 2016).

The key caveat is that **a predicted motif is not the same as a modified site**. A motif says that a residue *could* be modified by an enzyme with that preference. Whether it is modified in a given cell depends on whether the protein meets the enzyme: the right compartment, the right expression time, the right docking sites, and a residue that is accessible on the surface or in a disordered region. For example, many N-X-S/T sequons in secreted proteins are not glycosylated, and the topology of the protein constrains which can be (Zielinska et al. 2010). Motifs are short, so most proteins contain many matches by chance, and precision on a proteome scale is low.

Most training data now come from **mass spectrometry**. Phosphoproteomics has mapped hundreds of thousands of sites (Section 8.5), collected in databases such as PhosphoSitePlus (Hornbeck et al. 2014). These data have their own biases: abundant proteins and well-ionising peptides are over-represented, and a site that was not observed is not necessarily unmodified, so true negatives are lacking, as for protein interactions (Section 11.4). Moreover, being modified does not mean being functional: many observed phosphosites are poorly conserved and may be the result of non-specific kinase activity (Landry

et al. 2009). A machine-learning score that combines conservation, structure and other features has been used to rank the human phosphosites by their likely functional importance (Ochoa et al. 2019). For annotation, a predicted PTM site is a hypothesis, an observed site is a fact about modification, and a functional site needs further evidence, such as a phenotype of mutating the residue.

10.4.8 Disordered regions and short linear motifs

Not all function is carried by folded domains. Many proteins, especially in eukaryotes, contain long **intrinsically disordered regions** that do not fold on their own (Section 3.5) (Lee et al. 2014). They are rich in **short linear motifs** (SLiMs) (Kumar et al. 2021) of a few residues that bind folded domains of other proteins, carry modification sites, or act as degradation and targeting signals; the nuclear localisation signals and the KDEL motif in 3 are examples. Many regulatory and signalling functions depend on such regions. They are hard to annotate with the methods above: motifs are short, so they match by chance, they appear and disappear quickly in evolution, and disordered regions align poorly. Family- and domain-based annotation therefore says little about them, and a protein whose domains are all annotated may still have unannotated functions in its disordered regions.

10.4.9 Machine learning and language models

Most current methods combine several sources of evidence with machine learning. DeepGOPlus combines a convolutional network on the sequence with homology-based transfer (Kulmanov and Hoehndorf 2019). **Protein language models** (Section 1.10) provide embeddings that capture function beyond simple sequence similarity: annotations can be transferred from the protein that is nearest in embedding space rather than in sequence alignment, which works also for distant homologues (Littmann et al. 2021). For enzymes, contrastive learning on embeddings predicts EC numbers more accurately than alignment, especially for poorly characterised families (Yu et al. 2023). The best methods in recent CAFA rounds combine language models, homology, structure and networks (Section 10.7).

As for all supervised predictors, the reported accuracy of a function predictor depends on how the test proteins are related to the training proteins. If test proteins have close homologues among the training proteins, a method can score well by recognising the homologue rather than by learning anything general, and performance on new families is overestimated. Fair evaluations therefore split data by sequence similarity or by time (4), as CAFA does by design.

10.4.10 Combining evidence and the specificity of a claim

In practice, a function is assigned by combining several kinds of evidence: domains, characterised homologues, conserved functional residues, structure, localisation signals and genomic or network context. Two principles limit what can be concluded.

The first is that evidence must be **independent** to be counted twice. In Bayesian terms, the posterior odds for a function are the prior odds multiplied by a likelihood ratio for each piece of evidence, as in the functional association networks of Section 11.3, but only if the pieces are independent. They often are not. Two databases may display the same annotation because one copied it from the other, a network edge may be text-mined from the paper that supplied the GO term, and member databases of InterPro share seed sequences. Agreement among descendants of one annotation is *pseudo-replication*, and multiplying their confidences exaggerates the certainty (Figure 10.1).

The second is that an annotation should be no more **specific** than the evidence supports. Membership of an enzyme superfamily, a probable enzyme class (an EC number to three digits) and a specific substrate (four digits) are separate claims, each needing stronger evidence. **Rule-based annotation** makes the

required evidence explicit. A rule assigns a function only if, for example, a protein matches the family signatures, has the required active-site residues and a compatible domain architecture, belongs to an allowed taxonomic range, and has no excluding feature. UniProt uses such rules (UniRule and ARBA) for unreviewed entries (Section 2.3), and InterPro2GO maps a GO term to an InterPro entry only if the term holds for all members of the entry, so broad families get broad terms (Burge et al. 2012). Withholding an over-specific term is a correct decision, not a failure.

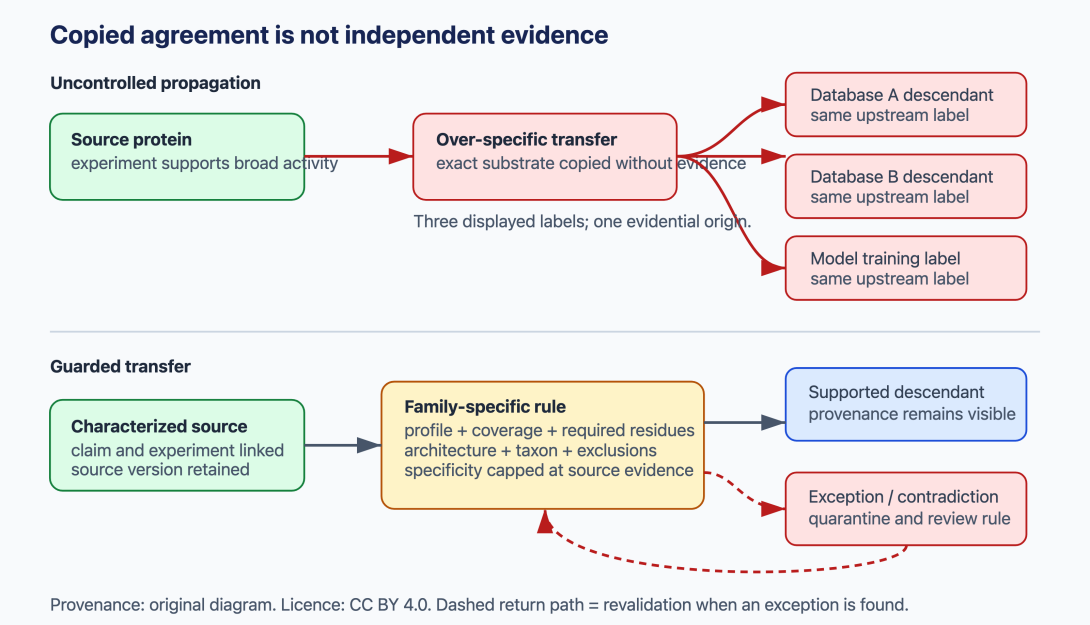


Figure 10.1: Copied agreement is not independent evidence. Above: an over-specific transfer (an exact substrate copied without evidence) spreads to several databases and to the training data of a predictor, which then appear to agree although they share one origin. Below: a family-specific rule that checks the profile, coverage, required residues, architecture and taxon, caps the specificity at that of the source evidence, keeps the provenance visible, and is revised when an exception is found. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

See also: further explanation in Section D.10.2; related Wikipedia articles in Section E.10.2.

Sources

Parts of this section are adapted from the Wikipedia articles Protein function prediction, Sequence homology and Phylogenetic profiling (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Function and Annotation" (CC BY 4.0), used with permission.

10.5 Subcellular localisation

Where a protein is located constrains what it can do: an enzyme in the mitochondrial matrix, a receptor in the plasma membrane and a transcription factor in the nucleus act in different contexts. Localisation is also one of the aspects of function that can be predicted most directly from sequence, because proteins are sorted to their destinations by signals in their own sequence.

10.5.1 Sorting signals

Günter Blobel proposed the **signal hypothesis** in the 1970s: proteins destined for secretion carry an N-terminal **signal peptide** that directs the ribosome to the endoplasmic reticulum, where the protein is translocated across the membrane and the signal is cleaved off (Blobel and Dobberstein 1975). Blobel received the 1999 Nobel Prize in Physiology or Medicine. Many kinds of sorting signal are now known:

- **Secretory signal peptides**, 15–30 residues at the N-terminus, with a positively charged *n-region*, a hydrophobic *h-region* and a polar *c-region* ending in a cleavage site recognised by signal peptidase, with small residues at positions −3 and −1 (the "−3, −1 rule") (Heijne 1986). Similar signals exist in bacteria, for export across the cytoplasmic membrane, including lipoprotein signals and twin-arginine (Tat) signals.
- **Transit peptides** at the N-terminus target proteins to mitochondria (amphipathic helices rich in positive residues), chloroplasts and the thylakoid lumen.
- **Nuclear localisation signals** (short basic motifs) and nuclear export signals, which can lie anywhere in the sequence.
- **Peroxisomal targeting signals**, such as the C-terminal tripeptide SKL (PTS1).
- **Retention and anchoring signals**, such as the C-terminal KDEL motif that keeps proteins in the ER, and signals for the attachment of GPI anchors.
- **Transmembrane segments**, which anchor proteins in membranes and are predicted as described in Section 9.10.

10.5.2 Predicting signal peptides

Signal peptides were first predicted with weight matrices around the cleavage site (Heijne 1986). **SignalP**, developed in Copenhagen, introduced neural networks (Nielsen et al. 1997), and later versions largely solved the main source of errors: hydrophobic **transmembrane helices** near the N-terminus resemble the h-region of signal peptides, so early predictors predicted many membrane proteins as secreted (Petersen et al. 2011). SignalP 6 uses a protein language model and predicts all five types of signal peptide in all domains of life, including their regions and cleavage sites (Teufel et al. 2022). Its predictions are among the most accurate in bioinformatics, because the signal is compact, well defined and supported by many experimentally verified examples.

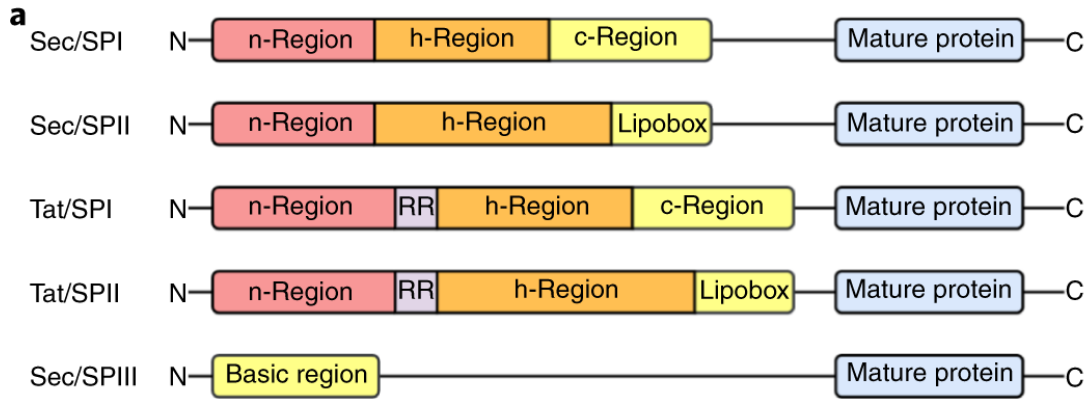


Figure 10.2: The five types of signal peptide predicted by SignalP 6.0: Sec/SPI (standard secretory signal peptides, cleaved by signal peptidase I), Sec/SPII (lipoprotein signal peptides with a lipobox, cleaved by signal peptidase II), the corresponding twin-arginine (RR) Tat/SPI and Tat/SPII signals, and Sec/SPIII (prepilin-like signals with a basic region). Figure 1a from Teufel et al. (2022) (CC BY 4.0), cropped by Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

10.5.3 Evaluating signal peptide and topology predictors

Signal peptide and transmembrane topology predictors label each residue, and their evaluation has some pitfalls of its own. The task can be posed in different ways: does the protein have a signal peptide (one label per sequence), where is the cleavage site (one position), or which residues belong to the n-, h- and c-regions (one label per residue). These are not interchangeable: a predictor can recognise a signal peptide without placing the cleavage site correctly, and residue-level accuracy is dominated by the many residues outside the signal. Performance should therefore be reported at the sequence level, as cleavage-site accuracy within a stated tolerance, and separately for each organism group and signal type. As for all predictors, training and test sets must be separated by homology, not randomly (4). Agreement between several predictors is not an experimental confirmation when they share training data, and a cleavage site observed by N-terminal proteomics is stronger evidence than any prediction.

10.5.4 Predicting localisation

Predicting the final location of a protein is harder, since it depends on signals anywhere in the sequence, on interactions and on regulation. Early programs such as PSORT encoded known sorting rules as an expert system (Nakai and Horton 1999). TargetP used neural networks to predict N-terminal targeting peptides for the secretory pathway, mitochondria and chloroplasts (Emanuelsson et al. 2000), and TargetP 2.0 used deep learning with attention to show which residues the network relied on (Almagro Armenteros et al. 2019). **DeepLoc** predicted ten locations from the whole sequence with a recurrent neural network with attention (Almagro Armenteros et al. 2017), and **DeepLoc 2** uses protein language model embeddings and predicts multiple locations for each protein, since many proteins are found in more than one compartment (Thumuluri et al. 2022). Training data come from UniProt annotations and, for human proteins, from imaging in the Human Protein Atlas, which showed that about half of all proteins are found in more than one location (Thul et al. 2017) (Section 8.6).

Localisation predictors are useful for annotating proteomes, for finding secreted proteins and drug targets, and for interpreting proteomics data. Their limitations are those of the training data: proteins that move between compartments, are targeted by unusual signals, or have locations that depend on the cell type are predicted poorly. How localisation is measured and classified from images is described in Section 13.5.

See also: further explanation in Section D.10.3; related Wikipedia articles in Section E.10.3.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Protein targeting, Signal peptide, Nuclear localization sequence and Protein subcellular localization prediction (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Signals, Membranes, and Localization" (CC BY 4.0), used with permission.

10.6 Predicting the effects of mutations

Every human genome contains about 10,000 missense variants, which change one amino acid in a protein. Most are harmless, a few cause disease, and for most of those found in patients the effect is unknown. Such *variants of uncertain significance* are one of the main obstacles in clinical genetics. **Variant effect prediction** estimates whether an amino acid substitution damages the function of a protein. It is function prediction at the level of single residues.

Three different properties are often mixed up under this heading. **Evolutionary constraint** describes whether substitutions like this one are tolerated at the position over evolution. A **molecular effect** is a change in the stability, activity or binding of the protein, as measured in an assay. **Pathogenicity** is a clinical label: whether the variant causes disease in a person. They are related but not the same. A variant can change the molecular function of a protein without causing disease, for example in a gene where one working copy is enough, and whether a damaging variant causes disease depends on the phenotype, the mode of inheritance, penetrance and the population. Most predictors are trained on, or evaluated against, only one of these, and their scores should be read accordingly.

10.6.1 Conservation-based predictors

The basic principle is evolutionary: positions that are conserved across homologues are functionally constrained, so substitutions there are likely to be harmful, while substitutions to amino acids seen at the same position in other species are likely to be tolerated.

- **SIFT** predicts whether a substitution is tolerated from the amino acid frequencies at the position in a multiple alignment of homologues (Ng 2003).
- **PolyPhen-2** combines conservation with structural features, such as burial and proximity to functional sites, in a naive Bayes classifier trained on known disease mutations (Adzhubei et al. 2010).
- **CADD** scores all possible variants in the genome, coding and non-coding, with a model trained to separate variants that have survived natural selection from simulated new mutations, using dozens of annotations (Kircher et al. 2014).

Supervised predictors trained on databases of known pathogenic and benign variants, such as ClinVar (Landrum et al. 2017), face a problem of **circularity**: the same variants, or variants from the same genes, are often used for training and for evaluation, and some clinical annotations were themselves based on earlier predictions. This inflates the apparent performance (Grimm et al. 2015). It is a form of the data leakage discussed in 4, and evaluations should keep genes and their homologues, not only individual variants, apart between training and test sets.

10.6.2 Unsupervised models of evolutionary sequences

A second generation of methods avoids labelled training data altogether. They learn the distribution of sequences in a protein family and score a variant by how much less likely the mutated sequence is than the wild type, under the model.

- **EVmutation** uses the same Potts model as coevolution-based structure prediction (Section 9.8), so that the effect of a mutation can depend on the rest of the sequence (*epistasis*) (Hopf et al. 2017).
- **DeepSequence** replaced the Potts model with a variational autoencoder (Riesselman et al. 2018), and **EVE** applied this approach to human disease genes, classifying variants without using any clinical labels, with accuracy comparable to supervised methods (Frazer et al. 2021).
- **Protein language models**, trained across all proteins rather than one family at a time, predict variant effects from the likelihood of the substituted amino acid at a masked position. ESM-1v (Meier et al. 2021) and ESM-1b, which was used to score all possible missense variants in the human proteome (Brandes et al. 2023), work even for proteins with few homologues.
- **AlphaMissense** fine-tuned an AlphaFold-based model, combining structural context and evolutionary information, on population frequency data as weak labels: variants common in humans and primates were treated as benign. It provides predictions for all 71 million possible human missense variants, and classifies most of them as likely benign or likely pathogenic (Cheng et al. 2023).

10.6.3 Experimental benchmarks

Deep mutational scanning measures the functional effect of thousands of variants of a protein in a single experiment, by linking a library of variants to a selection for function and counting each variant by sequencing before and after selection (Fowler and Fields 2014). These data provide benchmarks that are independent of clinical annotation, and unsupervised predictors correlate surprisingly well with them. Population data are a complementary source: variants that are common in healthy people are unlikely to cause severe disease, and genes in which damaging variants are rarer than expected are intolerant to mutation (Chen et al. 2023).

10.6.4 Using predictions

Clinical guidelines for interpreting variants allow computational predictions to be used only as supporting evidence, together with population frequency, segregation in families and functional data (Richards et al. 2015). Predictions are probabilities, not diagnoses, and they estimate whether a variant disrupts protein function, not whether it causes a particular disease: a damaging variant in a gene that tolerates loss of one copy may be harmless.

Most predictors also do not identify the **mechanism**. They largely measure constraint, or how deleterious a substitution is, and a high score does not say whether the variant causes **loss of function** (for example by destabilising the protein or removing a catalytic residue) or **gain of function** (for example mutations that make a kinase constitutively active, or the activating KRAS G12 mutations in cancer), or acts as a dominant negative. The distinction matters for interpretation: genes can cause disease through one mechanism and not the other, and gain-of-function variants can occur at positions where substitutions are not destabilising, so they may be scored as damaging for the wrong reason or be missed.

The Critical Assessment of Genome Interpretation (CAGI) evaluates these methods in blind challenges, much as CASP does for structure (The Critical Assessment of Genome Interpretation Consortium 2024).

See also: further explanation in Section D.10.4; related Wikipedia articles in Section E.10.4.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Missense mutation, Multiplexed Assays of Variant Effect and Variant of uncertain significance (CC BY-SA 4.0, retrieved 4 October 2026).

10.7 Benchmarking function prediction

How good are function predictions? Answering this is harder than for structure prediction, for reasons that are instructive for any field of bioinformatics.

10.7.1 CAFA

The **Critical Assessment of Functional Annotation** (CAFA) evaluates predictions **prospectively**, in the same spirit as CASP (Section 9.2). Organisers release the sequences of tens of thousands of proteins that lack experimental annotations. Participants predict GO terms for them, and after a waiting period of several months to years the predictions are evaluated on the proteins that have gained experimental

annotations in the meantime (Radivojac et al. 2013). Five rounds had been completed by 2024. CAFA2 and CAFA3 showed that methods improved, but that simple baselines, such as transferring annotations from the best BLAST hit or predicting the most frequent terms, were hard to beat by a large margin, particularly for biological process and cellular component (Jiang et al. 2016; Zhou et al. 2019). CAFA5 was run as a Kaggle competition, which brought in a large community of data scientists. Its methods improved on all methods from the previous rounds, helped by protein language models, predicted structures and the growth of curated experimental annotations. CAFA5 also introduced evaluation on proteins that already had some (incomplete) annotations, and showed that methods should make better use of existing annotations to predict new ones (De Paolis Kaluza et al. 2026).

Predictions are scored as precision and recall over the GO graph, summarised by the maximum F-measure over all score thresholds ($F_{\max}$). Before scoring, both the predicted and the experimental terms are propagated to their ancestors, so that the hierarchy is taken into account: a prediction of the parent of a correct term counts as partly correct rather than wrong. Because general terms are easy to predict and say little, measures that weight terms by their **information content**, such as the minimum semantic distance $S_{\min}$, reward predictions of specific terms (Clark and Radivojac 2013).

10.7.2 The CAFA measures

For protein i with experimental terms T_i and predicted terms $P_i(\tau)$ at score threshold τ, both propagated to their ancestors, the protein-centric precision and recall are

$$\mathrm{pr}(\tau) = \frac{1}{m(\tau)} \sum_{i} \frac{|P_i(\tau) \cap T_i|}{|P_i(\tau)|}, \qquad \mathrm{rc}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \frac{|P_i(\tau) \cap T_i|}{|T_i|},$$

where precision is averaged over the $m(\tau)$ proteins with at least one prediction and recall over all n proteins, and

$$F_{\max} = \max_{\tau} \frac{2\,\mathrm{pr}(\tau)\,\mathrm{rc}(\tau)}{\mathrm{pr}(\tau) + \mathrm{rc}(\tau)}.$$

$S_{\min}$ weights each term t by its **information content**, derived from its probability given its parents in the annotated corpus, so that specific, rare terms count more than general ones. The *remaining uncertainty* $ru(\tau)$ sums the information of true terms that were missed, the *misinformation* $mi(\tau)$ that of wrongly predicted terms, and $S_{\min} = \min_{\tau} \sqrt{ru(\tau)^2 + mi(\tau)^2}$, where lower is better (Clark and Radivojac 2013). Both measures depend on how proteins without predictions are handled, on whether averages are taken over proteins or over terms, and on the ontology release and evidence codes used.

10.7.3 Why function prediction is hard to evaluate

- **Incomplete annotations.** A protein's experimental annotations are never complete. A prediction that is not (yet) annotated is counted as a false positive, even if it is correct. This is the *open-world* problem. Measured precision is therefore usually an underestimate, and methods that predict plausible but not yet tested functions are penalised. Evaluation measures should be interpreted with this in mind; none fully corrects for it.
- **Multifunctionality and context.** Many proteins have several functions, some only in particular cell types, conditions or complexes (Section 10.2). Annotations usually record the context that was studied, so a predictor can appear wrong only because the database lacks the annotation for another context.

- **No negative data.** Databases record what proteins do, rarely what they do not do, so true negatives are hardly ever known.
- **Biased annotations.** Experimental annotations come mostly from a few organisms and a few high-throughput experiments, which favour general terms and particular functions (Schnoes et al. 2013). Evaluation on such data rewards methods that predict the common terms.
- **Changing ontology and circularity.** GO and the annotations change over time, and predictions that were deposited in databases can later be used, directly or indirectly, as evidence for new annotations.
- **The definition problem.** Since function has several meanings (Section 10.2), different evaluations measure different things, and a high score on GO terms does not guarantee that a prediction is useful to an experimentalist.

These problems explain why progress in function prediction looks modest compared with structure prediction, which has a single, well-defined and experimentally measurable answer. They also illustrate general lessons for evaluating predictors that are discussed in Chapter 12.

See also: further explanation in Section D.10.5; related Wikipedia articles in Section E.10.5.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Critical Assessment of Function Annotation and Protein function prediction (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Protein Function and Annotation" (CC BY 4.0), used with permission.

10.8 Key concepts

A checklist for revision.

- **What is function?** Biochemical activity, process, location, phenotype; context dependence; moonlighting and promiscuity
- **Gene Ontology:** molecular function, biological process and cellular component as separate annotation dimensions; the DAG and the true path rule (propagation over *is a* and *part of*); specificity of terms; evidence codes (experimental, phylogenetic, IEA); how few proteins have experimental annotations; what different kinds of evidence (assay, phenotype, interaction, expression, computation) show; annotation bias
- **GO annotations as statements:** relations (enables, involved in, located in); evidence and reference; the NOT qualifier and why it does not propagate upwards; annotation extensions; the open-world assumption; versions; GO-CAM causal models
- **EC numbers:** the four levels and seven classes; EC classifies reactions, not proteins; analogous enzymes; BRENDA and Rhea
- **Experimental determination of function:** forward vs. reverse genetics; mapping causal mutations by sequencing; the barcoded yeast deletion collection and pooled growth; IMPC; RNAi (knockdown, seed-mediated off-target effects) vs. CRISPR knockout, CRISPRi and CRISPRa; why knockdown and knockout phenotypes can disagree (residual protein, genetic compensation, off-targets); pooled CRISPR screens (guide counts, log fold changes, gene-level scores, MAGeCK, reference essential genes); the copy-number artefact and CERES/Chronos; core vs. context-specific essential genes; DepMap and co-essentiality; arrayed, imaging and Perturb-seq screens; redundancy, condition dependence, small effects and pleiotropy; a phenotype is not a molecular function

- **Annotation transfer by homology:** identity thresholds for enzymes as empirical tendencies, not rules; domain architecture; orthologues vs. paralogues and the ortholog conjecture (useful evidence, not a guarantee); eggNOG-mapper; error propagation and misannotation
- **Family- and domain-based annotation:** InterPro and InterPro2GO; curated annotation rules; FunFams; deep-learning family classification
- **Genomic context:** gene fusion (Rosetta stone), conserved gene neighbourhood, phylogenetic profiles; predicting links rather than functions
- **Networks:** guilt by association; propagation over networks
- **Structure-based prediction:** active-site templates; structural similarity; graph neural networks; proteome-wide annotation with predicted structures
- **Functional sites:** motifs (PROSITE), conservation and specificity-determining positions, pocket detection, catalytic-site templates; PTM site prediction (NetPhos, kinase-specific predictors, NetNGlyc/NetOGlyc and the N-X-S/T sequon, lipidation); a motif is not a modified site, and a modified site is not necessarily functional; MS data as training data and its biases
- **Disorder and short linear motifs:** functions not carried by domains; why they are hard to annotate
- **Combining evidence:** likelihood ratios and the independence assumption; pseudo-replication of copied annotations; the specificity of a claim (superfamily vs. EC class vs. substrate); rule-based annotation (UniRule, ARBA, InterPro2GO)
- **Machine learning:** DeepGOPlus; embedding-based annotation transfer; contrastive learning for EC numbers; homology between training and test sets
- **Localisation:** the signal hypothesis; signal peptides and their regions; transit peptides, NLS, PTS1, KDEL, GPI anchors; SignalP and the transmembrane-helix confusion; PSORT, TargetP, DeepLoc; multiple locations; the five SignalP 6 signal types (Sec/SPI–III, Tat/SPI–II); sequence-, cleavage-site- and residue-level evaluation; homology-aware splits
- **Variant effect prediction:** constraint vs. molecular effect vs. pathogenicity; conservation (SIFT), structure (PolyPhen-2), genome-wide scores (CADD); circularity of supervised methods; unsupervised family models (EVmutation, DeepSequence, EVE); protein language models (ESM-1v, ESM-1b); AlphaMissense; deep mutational scanning; gene constraint; clinical guidelines; loss vs. gain of function; CAGI
- **Benchmarking:** CAFA's prospective design; baselines; $F_{\max}$ (protein-centric precision and recall) and $S_{\min}$ (information content, remaining uncertainty and misinformation); propagation over the GO graph; information content; incomplete annotations and the open-world problem; multifunctionality and context; missing negatives; annotation bias; why function is harder to evaluate than structure

10.9 Recommended reading

These papers are recommended for reading in full:

- Ashburner et al. (2000) — the Gene Ontology; how function was turned into a computable vocabulary.
- Whisstock and Lesk (2003) — prediction of protein function from sequence and structure; the classical review.
- Marcotte, Pellegrini, Ng, et al. (1999) — the Rosetta stone method; predicting function from genome context rather than homology.
- Teufel et al. (2022) — SignalP 6.0; signal peptides of all types predicted with a protein language model.
- Cheng et al. (2023) — AlphaMissense; proteome-wide prediction of missense variant effects.
- Zhou et al. (2019) — CAFA3; how function prediction is evaluated, and what the results show.

Titles, abstracts, links and the full texts are in Section A.10; a longer list of further papers is in Section B.10.

10.10 Background reading (open access)

- **Protein function prediction (Wikipedia)** — Protein function prediction methods are techniques that bioinformatics researchers use to assign biological or biochemical roles to proteins. These proteins are usually ones that are poorly studied or predicted based on genomic sequence data.
  https://en.wikipedia.org/wiki/Protein_function_prediction
- **Gene Ontology (Wikipedia)** — The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
  https://en.wikipedia.org/wiki/Gene_Ontology
- **Enzyme Commission number (Wikipedia)** — The Enzyme Commission number is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. As a system of enzyme nomenclature, every EC number is associated with a recommended name for the corresponding enzyme-catalyzed reaction.
  https://en.wikipedia.org/wiki/Enzyme_Commission_number
- **Phylogenetic profiling (Wikipedia)** — Phylogenetic profiling is a bioinformatics technique in which the joint presence or joint absence of two traits across large numbers of species is used to infer a meaningful biological connection, such as involvement of two different proteins in the same biological pathway.
  https://en.wikipedia.org/wiki/Phylogenetic_profiling
- **Signal peptide (Wikipedia)** — A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined toward the secretory pathway. These proteins include those that reside either inside certain organelles, secreted from the cell, or inserted into most cellular membranes.
  https://en.wikipedia.org/wiki/Signal_peptide
- **Protein subcellular localization prediction (Wikipedia)** — Protein subcellular localization prediction involves the prediction of where a protein resides in a cell, its subcellular localization.
  https://en.wikipedia.org/wiki/Protein_subcellular_localization_prediction
- **Multiplexed Assays of Variant Effect (Wikipedia)** — Multiplexed Assays of Variant Effect (MAVE) is a method used to determine sequence-function relationships, identify variants that alter protein structure, stability, or activity, and characterize regulatory elements.
  https://en.wikipedia.org/wiki/Multiplexed_Assays_of_Variant_Effect

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

10.11 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

10.12 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters “Protein

Function and Annotation" and "Protein Signals, Membranes, and Localization" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

10.12.1 Decomposing protein function

A deletion strain grows poorly during oxidative stress, the encoded protein binds NADPH in vitro, and its transcript increases after peroxide treatment. Formulate the distinct functional claims supported by these observations. Which mechanistic claims remain unresolved?

Extensions, solution outline and common errors

Extensions:

- Distinguish molecular activity from biological process, localisation, phenotype and context.
- How could multifunctionality, isoforms, modifications or complex formation change the interpretation?
- Propose one experiment that connects NADPH binding to the stress phenotype.

Solution outline:

1. The deletion supports involvement of the gene in fitness under the tested condition (a mutant phenotype, IMP), subject to genetic background and indirect effects.
2. Induction of the transcript shows condition-dependent regulation (IEP), not protein activity.
3. NADPH binding supports a molecular interaction under the assay conditions (IDA), but not catalysis, physiological affinity or cellular relevance.
4. To connect them: test a chemically defined reaction and identify its products, compare the wild type with a folded variant that has lost the binding or catalytic site, and test whether the wild-type but not the variant rescues the deletion.
5. Represent function as separate assertions, each with its context and evidence, rather than as one name.

Common errors:

- Treating expression as mechanism.
- Inferring catalysis from binding.
- Assigning a pathway from a growth phenotype alone.
- Treating missing annotations as evidence of absence.

Connections: Section 10.2; ontologies and evidence codes in Section 2.6; expression in Chapter 7; interactions in Chapter 11.

10.12.2 Interpreting a domain match

A curated profile identifies a protein-kinase-like domain between residues 410 and 690 of a 1,200-residue protein. State the strongest defensible annotation and design the next computational and experimental steps.

Extensions, solution outline and common errors

Extensions:

- Define domain, motif, site, repeat and family.
- Discuss coverage, domain boundaries, pseudokinases, transmembrane segments, disorder and regulatory regions.
- Why may hits in several databases not be independent evidence?

Solution outline:

1. The match supports membership of residues 410–690 in the modelled domain family, given the score, coverage and sequence quality. It does not determine the function of the full-length protein.
2. Map the other domains, targeting signals, transmembrane segments and disordered regions, and check whether the catalytic residues required for kinase activity are present (pseudokinases lack them).
3. Inspect a confidence-filtered structural model and place the domain in a phylogeny of the family; trace the evidence for characterised relatives.
4. Kinase activity requires an assay with a suitable substrate and product readout, controls without enzyme and with an inactive variant, and evidence that the construct is folded.

Common errors:

- Naming the whole protein from a local hit.
- Using a short motif as proof of catalysis.
- Neglecting errors in the gene model.
- Counting matches to related signature databases (for example several InterPro member databases) as replication.

Connections: Section 10.4; profile searches and domains in Section 1.6 and Section 1.9; gene models in Section 5.6; predicted structures in Section 9.8.

10.12.3 Sequence identity and family membership

Two proteins share 45% identity over 220 aligned residues, but one is 780 residues long and the other 260. What can be concluded about ancestry, family membership and functional equivalence?

Extensions, solution outline and common errors

Extensions:

- Reconsider the case at 80% identity.
- Define a protein family at an explicit evolutionary depth.
- What evidence is required to transfer an exact enzyme specificity?
- How do low-complexity regions or biased composition affect the conclusion?

Solution outline:

1. Identity is a property of an alignment, not a biological classification. The aligned region is very probably homologous.

2. Incomplete coverage of the longer protein and different domain architectures rule out automatic transfer of the full-length family or function.
3. Family membership should be defined by common ancestry at a stated level, and supported by profiles or phylogeny.
4. Exact specificity also requires an experimentally characterised source, a compatible subfamily history, conserved catalytic and specificity-determining residues, and preferably biochemical evidence. Even high identity does not guarantee the same EC number, and no fixed percentage supplies these conditions (Rost 2002).

Common errors:

- Describing proteins as "45% homologous".
- Defining ancestry by a global identity cutoff.
- Ignoring the length over which identity is computed.
- Reading an E-value as the probability of equal function.

Connections: Section 10.4; homology in Section 3.3; alignment statistics in 1; domain evolution in Section 3.5.

10.12.4 Orthology and functional conservation

A human and a mouse protein are inferred to be one-to-one orthologues. Which functional properties can be transferred, and how would you test conservation?

Extensions, solution outline and common errors

Extensions:

- Why does orthology raise the prior probability of conserved function without guaranteeing it (the ortholog conjecture)?
- Consider isoforms, lineage-specific partners, expression, localisation and gene duplication.
- How does taxon sampling change an orthology call?

Solution outline:

1. Orthology means descent through speciation; it supports, but does not prove, conservation of function.
2. Evaluate a domain-aware alignment and gene tree reconciled with the species tree, synteny where informative, correspondence of isoforms, catalytic and regulatory residues, expression, localisation and interaction partners.
3. A biochemical activity shown in one species transfers more readily than tissue-specific regulation or a phenotype.
4. Each transferred assertion should keep its source and evidence code. Cross-species complementation supports conservation if expression, localisation and dosage are controlled.

Common errors:

- Inferring orthology only from a best reciprocal hit.
- Treating one-to-one status as independent of taxon sampling.
- Equating conserved chemistry with a conserved organismal phenotype.
- Counting several orthology resources built from the same data as independent.

Connections: Section 10.4; orthologues and paralogues in Section 3.4; gene-tree reconciliation in Section 4.4; expression in Chapter 7.

10.12.5 Reading a GO assertion

A record states that a protein *enables* protein kinase activity, with evidence code IEA. Interpret the assertion, and list the additional fields and versions needed to use it reproducibly.

Extensions, solution outline and common errors

Extensions:

- Contrast IEA, EXP, ND and an explicit NOT annotation.
- Explain propagation over the ontology (the true path rule) and the open-world assumption.
- Why does annotation to a broad ancestor not imply a specific descendant?

Solution outline:

1. IEA marks an automatically generated assertion that no curator has reviewed individually; it is not necessarily wrong.
2. Interpretation requires the gene product (and isoform), the relation, the GO identifier and ontology release, the reference, any qualifier or extension, the annotating group, the date and the pipeline.
3. Positive annotations propagate to ancestors only through relations such as *is a* and *part of*.
4. NOT is an evidence-based negation; ND records that curators found no biological data for that aspect. A missing annotation is unknown, not a negative (The Gene Ontology Consortium et al. 2023).

Common errors:

- Ranking evidence codes as universal truth scores.
- Ignoring the relation.
- Treating inherited ancestor terms as independent observations.
- Using unannotated proteins as negatives for machine learning.

Connections: Section 10.2; GO as a resource in Section 2.6; hierarchical evaluation in Section 10.7.

10.12.6 Structure-based functional inference

A predicted structure gives a high-scoring Foldseek match to a biochemically characterised enzyme, although sequence searches are inconclusive. Formulate a calibrated annotation strategy.

Extensions, solution outline and common errors

Extensions:

- Distinguish global fold similarity, remote homology, similarity of local sites and equivalence of mechanism.
- Discuss model confidence, uncertain side chains, oligomerisation, ligands and convergent geometry.
- Specify a discriminating biochemical assay.

Solution outline:

1. Structure search generates a candidate relationship (van Kempen et al. 2023). Restrict the comparison to confidently predicted domains and inspect the alignment extent and topology.
2. Add profile and phylogenetic evidence, and compare the equivalence of catalytic residues, pocket geometry, cofactors and metals, and quaternary structure.
3. A common fold supports a structural class but does not by itself establish ancestry or function; the narrowest provisional claim may be a fold or reaction class.
4. Exact function requires a substrate panel, kinetic measurements and identification of products, with controls for folding and contamination.

Common errors:

- Treating a predicted model as an experimental structure.
- Trusting uncertain side-chain positions.
- Reading global similarity as equivalence of the pocket.
- Using one docking pose as validation.

Connections: Section 10.4; structure comparison in Section 9.6; model confidence and docking in Section 9.8 and Section 9.11.

10.12.7 Network evidence and biological context

An uncharacterised protein is strongly connected to a DNA-repair module in a functional association network. What does the network support, and how would you validate it?

Extensions, solution outline and common errors

Extensions:

- Distinguish physical interaction from functional association.
- Explain hub effects, mixed conditions, dependence on text mining, and study bias.
- Describe an evaluation of network-based prediction that avoids leakage.

Solution outline:

1. The network supports association with the module under the meaning of its edges; it does not establish direct binding, molecular activity or a repair mechanism.
2. Separate the evidence channels (experiments, co-expression, genomic neighbourhood, text mining, pathways, orthology) and trace shared sources.
3. Guilt by association assumes that connected nodes share functions, an assumption weakened by hubs, multifunctional proteins and mixed conditions.
4. Validate by perturbing the protein under DNA damage, measuring repair-specific molecular and phenotypic outcomes, testing rescue, and using orthogonal interaction or localisation assays.
5. Evaluate predictions by holding out whole proteins or later annotations, not random edges.

Common errors:

- Equating a STRING edge with physical contact.
- Multiplying dependent scores.
- Ignoring compartment and condition.
- Validating only with generic stress sensitivity.

Connections: Section 10.4; functional association networks in Section 11.3; physical interactions in Section 11.2; leakage in 4.

10.12.8 Annotation propagation and audit

The same specific enzyme name occurs on thousands of bacterial homologues in several databases. Design an audit and a conservative replacement pipeline.

Extensions, solution outline and common errors

Extensions:

- Explain pseudo-replication across databases, paralogy and errors in domain boundaries.
- What should a family-specific annotation rule contain?
- How should corrected descendant annotations be handled?

Solution outline:

1. Trace each displayed name to its earliest evidential source; copied descendants are one line of evidence, not many.
2. Reconstruct gene models, domain architectures, subfamilies, duplication history, required residues, taxonomic range and known exceptions.
3. A replacement rule can require profile matches, coverage, architecture, specificity-determining residues and allowed taxa, with explicit exclusions. Its output must not be more specific than the experimental evidence across the matched set.
4. Report separate confidence for family, reaction class, substrate, process and location, with provenance.
5. Evaluate on experimentally characterised proteins with homology- or time-separated splits; misannotation is common in automatically annotated superfamilies (Schnoes et al. 2009).

Common errors:

- Voting across dependent databases.
- Training and testing on near-identical homologues.
- Correcting one record while leaving its descendants unchanged.
- Treating terms missing from a benchmark as experimentally disproved.

Connections: Section 10.4 (error propagation, combining evidence); database provenance in Section 2.2; evaluation in Section 10.7 and 4.

10.12.9 Calibrating protein-function inference

An uncharacterised bacterial protein is weakly similar to a known antibiotic-resistance enzyme. Construct an evidence chain for family membership, molecular activity and resistance of the organism, and say where the conclusions must remain provisional.

Extensions, solution outline and common errors

Extensions:

- How do domain architecture and alignment coverage change the interpretation?

- Which experiments distinguish biochemical activity from physiological relevance?

Solution outline:

1. Verify the sequence and gene model, then examine local similarity, profile evidence, domain architecture, conserved residues, phylogenetic placement, genomic context, predicted localisation and structure.
2. Treat homology as shared ancestry, not identical function: a match confined to a generic domain supports a narrower conclusion than full-length, family-specific evidence.
3. Check whether the reference annotations come from direct assays or computational transfer (evidence codes).
4. Test molecular activity with purified protein, substrates, negative controls and mutation of catalytic residues. Test resistance by deletion, complementation and susceptibility measurements.
5. Report a predicted family or activity until the evidence supports a stronger statement.

Common errors:

- Treating structural resemblance as proof of the same reaction.
- Ignoring paralogy.
- Inferring an organismal phenotype from an in vitro assay alone.

Connections: Section 10.4 and Section 10.2; homology in Section 3.3; sequence search and domains in 1 and Section 1.9.

10.12.10 Interpreting a pooled CRISPR screen

A genome-wide CRISPR knockout screen with four guides per gene is run in a cancer cell line for three weeks. Gene A has all four guides strongly depleted; gene B has one guide strongly depleted and three unchanged; a cluster of ten neighbouring genes on an amplified region of chromosome 8 all appear depleted; gene C, a paralogue of gene D, shows no depletion. Which genes would you call dependencies, and what further analyses or experiments would you do?

Extensions, solution outline and common errors

Extensions:

- How would you decide whether gene A is a core essential gene or a dependency specific to this cell line?
- What would a CRISPRi screen add, and why might an RNAi screen of the same genes disagree?

Solution outline:

1. Normalise guide counts for sequencing depth, compute log fold changes against the starting population, and combine guides into gene-level scores with a robust method (for example MAGeCK's rank aggregation), controlling the false discovery rate. Check the screen's quality by how well it separates reference essential and non-essential genes.
2. Gene A: consistent depletion of independent guides is strong evidence of a fitness dependency.
3. Gene B: one depleted guide is more likely an off-target or guide-specific effect; it should not be called a hit without additional guides.

4. The amplified cluster: depletion of every gene in an amplified region is the copy-number artefact of Cas9 cutting at many sites. Correct with CERES or Chronos, or test with CRISPRi, which does not cut.
5. Gene C: no depletion does not mean no function. Paralogue redundancy with D is likely; test a double knockout for synthetic lethality, and other conditions.
6. Compare with DepMap: a gene essential in all lines is core essential; one essential only in lines with a particular mutation is a candidate selective target. Validate with independent guides and rescue by a guide-resistant cDNA.

Common errors:

- Calling hits from single guides.
- Ignoring copy number.
- Interpreting absence of depletion as absence of function.
- Equating a fitness phenotype with a molecular function.

Connections: Section 10.3; count statistics in Section 7.4; cancer genomics in Section 7.7; genetic interactions in Section 11.2.

11 Protein–protein interactions

> **From the exam description**
>
> Experimental and computational methods to analyse which proteins interact and how.

11.1 Overview

Proteins rarely act alone. The word *interaction*, however, is used for relationships of very different kinds, which are not interchangeable: direct **physical binding** between two proteins; **co-complex membership**, in which two proteins belong to the same complex without necessarily touching; **genetic interactions**, in which mutations in two genes have a combined effect on the phenotype; **regulatory influence**, in which one gene product changes the expression or activity of another; and **functional association**, in which two proteins take part in the same process. Each experimental and computational method detects one or a few of these. This chapter is broad: it covers physical interactions between proteins, functional associations, regulatory networks and the network view of the cell, as well as the prediction of interactions and of the structures of complexes:

- experimental methods: yeast two-hybrid, AP-MS and other methods, binding affinity and live-cell proximity, their errors and biases, hu.MAP, genetic interactions, and gene regulatory networks inferred from perturbations (Section 11.2);
- networks: interactome maps, network concepts and their critics, functional association networks and Bayesian integration, pathway and network enrichment analysis, disease networks, systems biology and flux balance analysis, and dynamic models of regulatory networks (feedback, switches, oscillators, robustness and parameter estimation) (Section 11.3);
- prediction methods: interfaces, genomic context, coevolution, docking and deep learning with AlphaFold, and how predictions are evaluated (Section 11.4).

The interaction databases are described in Section 2.7, the mass spectrometry methods in Section 8.5, and the structure prediction methods on which complex prediction builds in Chapter 9.

11.2 Experimental methods

Most proteins carry out their function together with others: in stable complexes such as the ribosome, in transient signalling interactions, or as enzymes acting on protein substrates. Mapping which proteins interact, and how, is a way to understand cellular machinery and to annotate unknown proteins (Section 10.4). The word *interaction* covers very different relationships, and each experimental method measures a different one. This section describes the main methods for physical interactions, the resulting maps and their errors, and the inference of regulatory networks from perturbations.

11.2.1 Binary interactions: yeast two-hybrid

The **yeast two-hybrid** (Y2H) system, invented by Stanley Fields in 1989 (Fields and Song 1989), exploits the modular structure of transcription factors. The DNA-binding domain of the yeast factor Gal4 is fused to one protein (the *bait*) and its activation domain to another (the *prey*). Only if bait and prey bind each other is a functional transcription factor reconstituted, which switches on a reporter gene that allows the yeast to grow on selective medium. Y2H tests one pair at a time and therefore enriches for **binary**, often direct, interactions, and can be run as large matrix screens of thousands of baits against thousands of preys. A positive result does not prove direct contact in the native cellular context, however: an endogenous yeast protein can bridge bait and prey, the fusion proteins are expressed outside their normal location and at unnatural levels, and the reporter can be activated without binding. The fusions can also misfold or hide the binding site, which causes false negatives.

The first proteome-scale screens were done in yeast (Uetz et al. 2000; Ito et al. 2001), and later for human proteins (Rual et al. 2005). The results were noisy: the two yeast screens overlapped little, and there was a widespread belief that Y2H data were unreliable. Marc Vidal's group showed that careful screening, retesting and an empirical framework for estimating precision and sensitivity gave high-quality data, and that the low overlap mostly reflected low *sensitivity*, since each screen samples only a fraction of all interactions (Yu et al. 2008; Braun et al. 2008; Venkatesan et al. 2008). The human reference interactome **HuRI** screened about 17,500 proteins against each other, and reported about 53,000 binary interactions (Rolland et al. 2014; Luck et al. 2020).

Y2H has its own biases. Interactions are tested in the yeast nucleus, so membrane proteins, interactions that require post-translational modifications or other partners, and proteins that activate transcription by themselves are problematic. Variants such as split-ubiquitin systems handle membrane proteins, and the **protein-fragment complementation assay** (PCA), in which two halves of a reporter enzyme are fused to the proteins and reassemble when they meet, tests interactions in their normal cellular location (Tarassov et al. 2008).

11.2.2 Co-complex interactions: affinity purification–mass spectrometry

In **affinity purification–mass spectrometry** (AP-MS), a tagged bait is purified from cells together with the proteins bound to it, which are identified by mass spectrometry (Section 8.5). The tandem affinity purification (TAP) tag, with two successive purification steps, made it possible to isolate complexes under near-native conditions (Rigaut et al. 1999), and was used for the first proteome-scale maps of yeast complexes (Gavin et al. 2002; Ho et al. 2002; Krogan et al. 2006). In human cells, the BioPlex project has purified more than 10,000 baits in two cell lines (Huttlin et al. 2015, 2021).

AP-MS reports **co-complex membership**: a prey can bind the bait directly or through other proteins. This is the key contrast with Y2H, which tests pairs and enriches for direct binary contacts, whereas AP-MS finds the proteins that co-purify with the bait in a cell, without telling which of them touch each other. A purified complex with n components can be represented in two ways, which give very different numbers of "interactions": the *spoke* model connects the bait to every prey ($n-1$ edges), while the *matrix* model connects all pairs ($n(n-1)/2$ edges), most of which are not direct contacts. Contaminants and abundant sticky proteins are removed with statistical scoring against control purifications (Section 8.5). Complementary MS methods, co-fractionation of native complexes, proximity labelling and cross-linking MS (O'Reilly and Rappsilber 2018), are also described in Section 8.5. They too measure different things: co-fractionation finds proteins that elute together, and proximity labelling finds proteins that were *near* the bait, within roughly 10–20 nm, not necessarily bound to it. **Cross-linking MS** is a bridge between interaction mapping and structure: it covalently links residues that are close in space, so each identified cross-link shows not only that two proteins are near each other but also which of their residues

are, typically within about 30 Å. These distance restraints can be used to dock or model the complex (Section 11.4), and can be collected in intact cells.

hu.MAP integrates thousands of AP-MS, co-fractionation and proximity labelling experiments with machine learning to predict human complexes (Drew et al. 2017, 2021; Fischer et al. 2025). How the map is built is described in Section 8.5. For network analysis, hu.MAP provides a set of complexes with confidence levels that complements the binary map of HuRI and the pull-downs of BioPlex.

11.2.3 Binding affinity, live-cell proximity and validation

Large-scale screens show whether two proteins associate, not how strongly, for how long or where. For a reversible 1:1 interaction $A + B \rightleftharpoons AB$, the dissociation constant is $K_d = [A][B]/[AB] = k_{\text{off}}/k_{\text{on}}$ (Section 9.3). If B is in excess and A has a single binding site, the fraction of A that is bound is

$$f_{\text{bound}} = \frac{[B]}{K_d + [B]},$$

so half of A is bound when the free concentration of B equals K_d. Whether an interaction occurs in a cell therefore depends on local concentrations as much as on affinity, and two interactions with the same K_d can have very different lifetimes ($1/k_{\text{off}}$). The simple equation assumes one independent site at equilibrium; cooperativity, multivalency (avidity) and competition between partners break it.

Affinity and kinetics are measured with purified proteins. **Surface plasmon resonance** and **biolayer interferometry** immobilise one partner and follow binding of the other in real time, which gives k_{on} and k_{off}; **isothermal titration calorimetry** measures the heat of binding, which gives K_d, the stoichiometry and the enthalpy. Such assays show direct binding under the assay conditions, but not that it happens in the cell.

In living cells, proximity can be measured with fluorescent tags. In **Förster resonance energy transfer** (FRET), an excited donor fluorophore transfers energy to a nearby acceptor with efficiency

$$E = \frac{1}{1 + (r/R_0)^6},$$

where r is the donor–acceptor distance and R_0, the Förster distance at which $E = 0.5$, is a few nanometres for common pairs (Stryer and Haugland 1967). The steep sixth-power dependence makes FRET sensitive only within a narrow range of distances. **BRET** uses a luciferase as donor, which avoids exciting the sample with light. A positive signal shows that the tags were close, which overexpression and molecular crowding can also cause, and a negative signal can reflect unfavourable tag geometry. Donor-only controls, non-binding mutants and titration of expression levels help to tell these apart. **Proximity labelling** (Section 8.5) records which proteins were near a bait during a labelling window rather than which bind it: hours for BioID (Roux et al. 2012), about ten minutes for TurboID (Branon et al. 2018), and about a minute for the peroxidase APEX, which requires hydrogen peroxide (Rhee et al. 2013).

Confidence grows when methods with *different failure modes* agree. Repeating Y2H with another complementation assay adds little, because both share the artefacts of fusion proteins; co-IP and AP-MS share the artefacts of cell lysis; and two databases may report the same original paper. A strong chain of evidence might combine discovery by AP-MS, reciprocal co-immunoprecipitation of the endogenous proteins, binding of the purified proteins, mutations of interface residues that abolish binding without unfolding either protein, and a functional defect that is rescued by the wild-type but not by the mutant protein. A single negative result rarely rules out an interaction, since every method has blind spots.

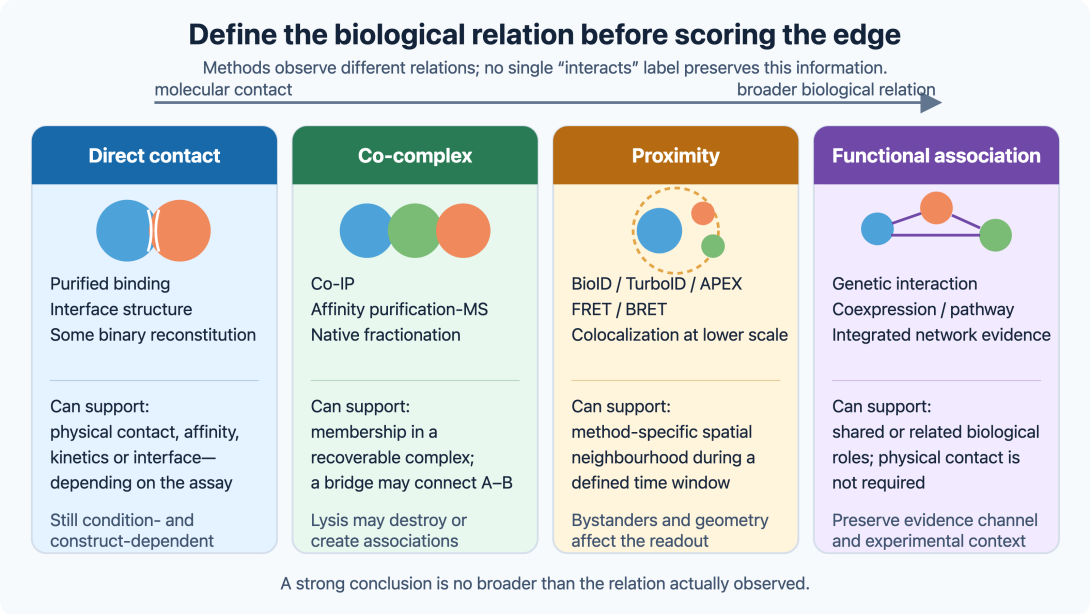


Figure 11.1: What different kinds of interaction evidence can support. Direct-contact methods (purified binding, interface structures) can show physical contact, affinity and kinetics; co-complex methods (co-IP, AP-MS, native fractionation) show membership in a recoverable complex, possibly bridged by other proteins; proximity methods (BioID, TurboID, APEX, FRET, BRET) show that proteins were near each other during a time window; and functional association (genetic interactions, co-expression, integrated networks) does not require contact at all. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

11.2.4 Errors, biases and coverage

Early comparisons of yeast interaction data sets found that they overlapped surprisingly little and estimated high rates of false positives and false negatives (Mering et al. 2002). Several lessons have emerged since then:

- **Different methods measure different things.** Binary Y2H interactions and co-complex AP-MS associations are not expected to agree, and both differ from functional associations in networks such as STRING (Section 11.3).
- **Low overlap often means low sensitivity, not low precision.** Each screen detects only part of the interactome, so two good screens can overlap little (Venkatesan et al. 2008).
- **Literature-curated data are biased.** Well-studied proteins, such as cancer genes, have many reported interactions simply because many experiments were done on them, which creates artificial hubs. Systematic screens, in which every pair is tested equally, avoid this *study bias* (Rolland et al. 2014; Luck et al. 2020).
- **Interactions depend on context.** Two proteins can only interact if they are expressed in the same cell, at the same time and in the same compartment, and many interactions depend on the isoform, on post-translational modifications or on a stimulus. There is therefore no single, fixed human interactome: each cell type and condition uses a different subset. A static network such as HuRI or BioPlex is a union of interactions observed or possible across many contexts, and many of its edges may never occur in a given cell.
- **Interactions are not binary.** Databases and networks record an interaction as present or absent, but biological interactions differ enormously in strength and dynamics. Dissociation constants (K_d) range from picomolar or tighter for stable complexes to micromolar or weaker for transient signalling interactions, residence times range from milliseconds to the lifetime of the protein,

and complexes have defined stoichiometries (for example two copies of one subunit and one of another). Whether an interaction occurs in the cell depends on K_d and on the concentrations of the partners, and which interactions an assay detects depends on its sensitivity to weak and transient binding.

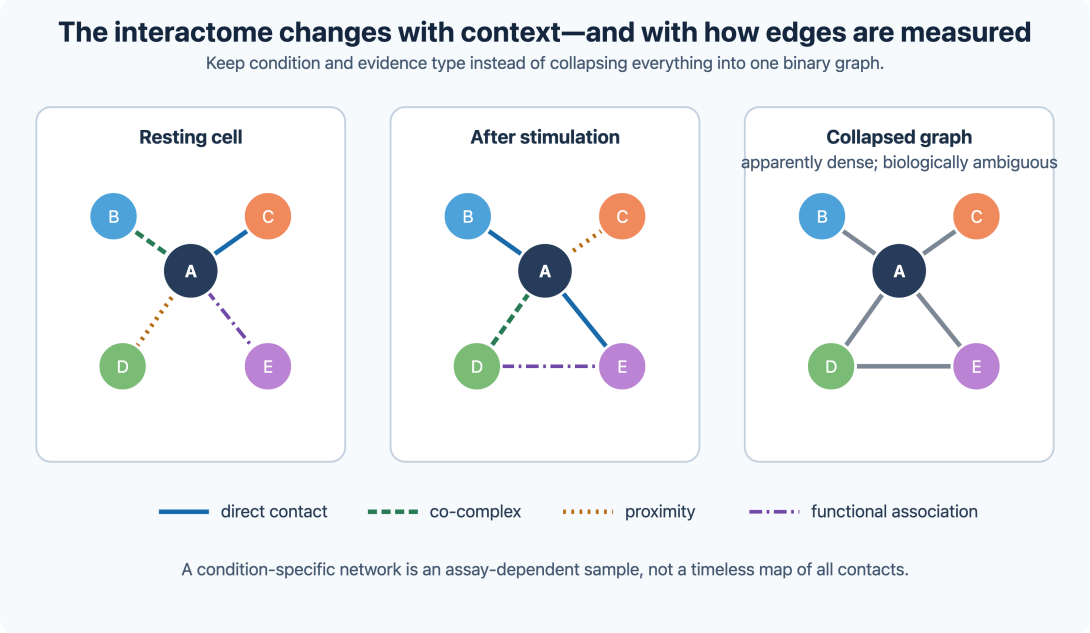


Figure 11.2: The same five proteins in a resting cell, after stimulation, and with all edges collapsed into one graph. Edges differ in kind (direct contact, co-complex, proximity or functional association) and in the condition in which they were seen; the collapsed graph looks denser and better connected than either real state. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

11.2.5 Genetic interactions

A **genetic interaction** occurs when the combined effect of mutations in two genes differs from what is expected from their individual effects. In *synthetic lethality*, two mutations that are harmless alone kill the cell together. Genetic interactions do not imply physical contact, but they reveal functional relationships: genes with similar profiles of genetic interactions tend to work in the same pathway or complex. A map of almost all pairwise genetic interactions in yeast showed a hierarchy of functional modules (Costanzo et al. 2016). Synthetic lethality is also a strategy for cancer therapy, targeting a gene whose partner is mutated in the tumour.

Quantitatively, a genetic interaction is a deviation from a null model. If w_A, w_B and w_{AB} are the fitness of the single and double mutants relative to the wild type, the usual multiplicative model expects $w_{AB} = w_A w_B$, and the interaction score is $\varepsilon_{AB} = w_{AB} - w_A w_B$. Negative scores (*aggravating* interactions, with synthetic lethality as the extreme case) often connect genes in parallel, buffering pathways; positive scores (*alleviating* interactions) often connect genes in the same pathway or complex, where losing one already disables the whole. An additive or logarithmic null model gives different scores, so which pairs "interact" depends on the model chosen.

11.2.6 Gene regulatory networks and inference from perturbations

A **gene regulatory network** (GRN) describes which transcription factors regulate which genes. Its edges are directed and signed (activation or repression), unlike the undirected edges of protein interaction networks. Regulatory interactions can be measured directly, for example by ChIP-seq of transcription factors, but they are often *inferred* from expression data.

- **Correlation and information-based methods** connect genes whose expression co-varies across many samples. ARACNE uses mutual information and removes indirect links with the data-processing inequality (Margolin et al. 2006). GENIE3 predicts the expression of each gene from all others with random forests and ranks the regulators by importance (Huynh-Thu et al. 2010).
- **Perturbation-based methods** use experiments in which genes are knocked down, knocked out or overexpressed. Since a perturbation is a cause, these data can reveal the direction of regulation, which correlations cannot. Early methods fitted linear dynamical models to the responses to known perturbations (Gardner et al. 2003; Tegnér et al. 2003), and the idea that perturbation data reveal the network structure goes back to the systematic perturbation of the yeast galactose pathway (Ideker et al. 2001).
- **Single-cell perturbation screens**, such as Perturb-seq, combine pooled CRISPR perturbations with single-cell RNA-seq, so that the transcriptional effect of each perturbation is read out in the same cells (Dixit et al. 2016). Genome-scale Perturb-seq has measured the effect of perturbing essentially every expressed gene in human cells (Replogle et al. 2022).

Network inference is notoriously difficult. The DREAM challenges compared dozens of methods on simulated and real data, and found that no single method was best, that performance on real data was modest, and that a consensus of several methods ("the wisdom of crowds") was most robust (Marbach et al. 2012). Benchmarks on single-cell data reached similar conclusions (Pratapa et al. 2020). Inferred networks should therefore be treated as hypotheses, and their accuracy depends strongly on how much of the network was actually perturbed.

See also: further explanation in Section D.11.1; related Wikipedia articles in Section E.11.1.

Sources

Parts of this section are adapted from the Wikipedia articles Protein–protein interaction, Two-hybrid screening, Tandem affinity purification, Synthetic lethality and Gene regulatory network (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Molecular Interactions: Experiments, Structures, and Networks" (CC BY 4.0), used with permission.

11.3 Networks

Interactions together form networks, and network science provides the concepts and tools to study them (Barabási and Oltvai 2004). This section describes the large interactome maps, the main properties of biological networks, functional association networks that integrate many kinds of evidence, and two applications: network medicine and metabolic modelling.

11.3.1 Interactome maps and databases

Three large, systematic maps of human protein interactions are the references for network analysis today:

- **HuRI**, the binary interactome from systematic yeast two-hybrid screens (Luck et al. 2020);
- **BioPlex**, the co-complex network from AP-MS in two cell lines (Huttlin et al. 2021);
- **hu.MAP**, protein complexes integrated from many MS experiments (Fischer et al. 2025).

These are complemented by curated databases of interactions reported in the literature, such as IntAct and BioGRID, and by functional association networks, such as STRING and FunCoup, which are described as resources in Section 2.7.

11.3.2 Network concepts

A network (graph) consists of **nodes** (proteins or genes) and **edges** (interactions). Edges may be undirected (physical binding), directed (regulation, signalling) or weighted (confidence). The main descriptive properties are:

- The **degree** k of a node is its number of edges, and the **degree distribution** $P(k)$ the fraction of nodes with degree k. In a random (Erdős–Rényi) graph, degrees follow a Poisson distribution around the mean.
- **Scale-free networks** have a power-law degree distribution, $P(k) \propto k^{-\gamma}$, so that most nodes have few links and a few **hubs** have very many. Barabási and Albert showed that such networks arise from *preferential attachment*: new nodes link preferentially to nodes that already have many links (Barabási and Albert 1999). In the early 2000s, metabolic and protein interaction networks were reported to be scale-free (Jeong et al. 2000), and gene duplication, in which a copied gene inherits the interactions of its parent, was proposed as a biological mechanism for preferential attachment. These claims are now regarded as controversial (see below).
- The **clustering coefficient** of a node is the fraction of pairs of its neighbours that are themselves connected: for a node i with k_i neighbours and e_i edges among them, $C_i = 2e_i / [k_i(k_i - 1)]$. Biological networks have high clustering and short paths between most nodes, the **small-world** property (Watts and Strogatz 1998).
- **Centrality** measures how important a node is for the network, for example by its degree or by the number of shortest paths passing through it. The *betweenness* of node i is $C_B(i) = \sum_{s \neq i \neq t} \sigma_{st}(i) / \sigma_{st}$, where σ_{st} is the number of shortest paths between nodes s and t and $\sigma_{st}(i)$ the number of those that pass through i. Centrality describes a position in one particular graph; it is not a measure of biological importance.
- **Modules** or **communities** are groups of nodes more densely connected to each other than to the rest, often corresponding to complexes or pathways. They are found with clustering algorithms such as Girvan–Newman edge removal (Girvan and Newman 2002) and the Markov cluster algorithm (Enright 2002). Many methods maximise the *modularity* $Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$, where A is the adjacency matrix, m the number of edges and $\delta(c_i, c_j) = 1$ if nodes i and j are in the same module: the number of edges within modules minus the number expected in a random network with the same degrees. A module found in this way is a hypothesis, not necessarily a complex. Metabolic networks show a hierarchical organisation of modules (Ravasz et al. 2002).
- **Network motifs** are small subgraphs, such as the feed-forward loop in transcriptional networks, that occur much more often than in randomised networks with the same degrees (Milo et al. 2002).

Early studies linked network position to biology. In yeast, hubs were found to be more often essential than other proteins, a correlation that became known as the *centrality–lethality rule* (Jeong et al. 2001).

Hubs were divided into *party hubs*, which interact with their partners at the same time, inside complexes, and *date hubs*, which bind different partners at different times (Han et al. 2004), and they differ in sequence and domain properties (Ekman et al. 2006). Interactions between hubs were found to be rarer than expected (Maslov and Sneppen 2002).

Many of these conclusions have been questioned. Interaction maps are incomplete and biased, so the observed network is a sample of the true one, and sampling can create or hide apparent power laws (Hakes et al. 2008; Lima-Mendez and Helden 2009). A heavy-tailed degree distribution is not necessarily a power law: other distributions, such as the log-normal or a power law with an exponential cut-off, often fit the data as well or better. A systematic analysis of nearly a thousand networks from many fields found that strongly scale-free networks are rare (Broido and Clauset 2019). Hubs in literature-curated networks partly reflect study bias (Section 11.2).

The same biases confound the centrality–lethality correlation. Essential proteins are studied more, and therefore accumulate more reported interactions. When AP-MS data are represented with the matrix model, every member of a large complex becomes a hub, and many large complexes are essential. The strength of the correlation depends on the data set and the detection method, and it does not show that a protein is essential *because* it has many partners (He and Zhang 2006). Network properties are therefore best treated as descriptions of a particular data set, and compared with appropriate randomised networks, rather than as universal laws.

11.3.3 Functional association networks

Physical interactions are only one kind of evidence that two proteins work together. Others are co-expression, genetic interactions, shared genomic context (Section 10.4), co-occurrence in the literature, shared domains and interactions of orthologues in other species. **Functional association networks** integrate all of these into a single network in which each edge has a confidence score (Guala et al. 2019).

The integration is usually **Bayesian**. For each evidence type i, a *likelihood ratio* is estimated from a gold standard of known positive pairs (for example proteins in the same complex or pathway) and negative pairs:

$$\mathrm{LR}_i = \frac{P(e_i \mid \text{linked})}{P(e_i \mid \text{not linked})}.$$

If the evidence types are assumed to be independent (*naive Bayes*), the posterior odds that two proteins are linked are the prior odds multiplied by the likelihood ratios of all the evidence, so a link supported by several independent weak signals can become confident. This approach was introduced for yeast (Jansen et al. 2003; Lee et al. 2004). **FunCoup**, developed by Erik Sonnhammer's group in Stockholm, applies it across many species and evidence types, and transfers evidence between species by orthology (Alexeyenko and Sonnhammer 2009; Buzzao, Persson, et al. 2024). **STRING** uses a similar scheme, with scores for each evidence channel combined into an overall confidence (Szklarczyk et al. 2024). If the channel scores s_i are treated as independent probabilities, the combined score is $1 - \prod_i (1 - s_i)$, the probability that at least one channel is right, computed after the prior probability that two random proteins are linked has been removed from each channel. The assumption of independence is not exact, for example because co-expression and shared pathway membership are correlated, and the networks reflect the biases of their gold standards.

An edge in such a network means that two proteins probably **work together**, not that they bind each other. Functional association networks are used to predict function by guilt by association (Section 10.4), to interpret gene lists from expression or genetic screens, and to prioritise disease genes.

11.3.4 Pathways, gene sets and network enrichment

Gene lists from expression, proteomics or genetic screens are usually interpreted with **gene sets**: GO terms and pathways from KEGG and Reactome (Section 2.6). Over-representation analysis with the hypergeometric test, gene set enrichment analysis (GSEA), and the choice of background and threshold are described in Section 7.5. Several further problems arise from the structure of the annotations and from the use of networks.

- **Dependence in the GO graph.** Annotations propagate to all ancestors of a term (Section 10.2), so a term and its parents share most of their genes. If a specific term is enriched, its ancestors usually are too, even when they carry no additional signal, and the result is a long, redundant list. The *elim* and *weight* algorithms of topGO remove or down-weight the genes of significant descendant terms before their parents are tested (Alexa et al. 2006), and *parent–child* analysis tests a term against the genes annotated to its parents rather than against the whole background (Grossmann et al. 2007). Long lists can also be summarised by clustering terms by semantic similarity, for example with REVIGO (Supek et al. 2011).
- **Multiple testing.** Thousands of overlapping sets are tested, so p-values must be corrected, usually by controlling the false discovery rate (Section 7.4). Because the tests are strongly dependent, the number of significant sets says little about the number of independent findings. Results also change with the database, its release and the identifier mapping, which should be reported (Wijesooriya et al. 2022; Chicco and Agapito 2022).
- **Annotation bias and circularity.** Well-studied and multifunctional genes belong to many sets and can drive both enrichment and guilt by association results on their own (Gillis and Pavlidis 2011). A network calibrated with pathway databases, as STRING partly is, cannot provide independent evidence for the same pathways.

Over-representation analysis and GSEA treat a pathway as an unordered set of genes. **Topology-based pathway analysis** also uses the directed, signed edges of a signalling pathway. SPIA, for example, combines the usual enrichment p-value with a perturbation score that propagates the measured changes along the pathway, so that a change in an upstream receptor counts for more than a change in a peripheral gene (Tarca et al. 2008).

Network enrichment analysis goes further. It asks whether the genes in a list have more links to the genes of a pathway, in a functional association network such as FunCoup, than expected by chance. A pathway can then be detected even if few of the query genes are annotated to it, which helps when annotations are incomplete. The null model is essential: hubs link to everything, and large pathways receive many links by chance, so the expected number of links is estimated from randomised networks that preserve the degree of each node (Alexeyenko et al. 2012), or with a binomial model of link counts, as in BinoX (Ogris et al. 2016). ANUBIX samples random gene sets against the intact pathway rather than treating pathways as random gene sets, which removes a bias towards particular pathways (Castresana-Aguirre and Sonnhammer 2020), and is available with other methods in the PathBIX web server (Castresana-Aguirre et al. 2021). In a benchmark that used related KEGG pathways as a gold standard for 82 disease expression data sets, network enrichment methods performed best, and most methods gave skewed p-values on randomised data (Buzzao, Castresana-Aguirre, et al. 2024); the benchmark came from the group that developed several of these methods. The same idea is used to prioritise genes: MaxLink ranks candidate genes by how many links they have to a seed set of known disease genes, compared with what their degree would predict (Guala et al. 2014).

Overlap-based and network-based methods answer different questions and can legitimately disagree. Network methods inherit the incompleteness and study bias of the network (Section 11.2), and are not automatically better.

11.3.5 Disease networks and network medicine

Genes associated with the same disease tend to interact with each other or lie close together in the network. A network of human diseases, connected when they share associated genes, showed that disease genes form clusters (Goh et al. 2007). **Network medicine** uses this to find new disease genes, to relate diseases to each other and to find drug targets (Barabási et al. 2010). Diseases correspond to network neighbourhoods (*disease modules*), and diseases whose modules overlap tend to share clinical features and comorbidities (Menche et al. 2015). Candidate genes from genetic studies are prioritised by their network proximity to known disease genes, for example with **random walks** or **network propagation**, in which signal diffuses from known genes through the network and nodes that receive much of it are ranked highly (Köhler et al. 2008; Vanunu et al. 2010). In a random walk with restart, a probability vector over the nodes is updated as $\mathbf{p}_{t+1} = (1-r)W\mathbf{p}_t + r\mathbf{p}_0$, where W is the column-normalised adjacency matrix, $\mathbf{p}_0$ puts all probability on the known (seed) genes and r is the probability of jumping back to them; the steady-state $\mathbf{p}$ ranks the candidates. Such methods are evaluated by hiding some of the known genes and testing whether they are ranked highly, which is only fair if the network was not itself built from the same disease annotations. A related idea is to look for *active subnetworks*, connected regions of the network in which many genes change in an experiment (Ideker et al. 2002). A network of genes in *C. elegans* could predict the phenotypes of perturbations (Lee et al. 2008).

11.3.6 Systems biology and metabolic modelling

Systems biology aims to understand how the behaviour of a system emerges from its components and their interactions, by combining large-scale measurements, perturbations and quantitative models (Kitano 2002; Ideker et al. 2001). Models range from qualitative networks to systems of ordinary differential equations describing the kinetics of each reaction, which require parameters that are rarely known.

Metabolism is the best-mapped network in the cell. You should know its outline: **glycolysis** breaks glucose down to pyruvate, the **citric acid cycle** oxidises acetyl-CoA to CO_2 and produces reduced electron carriers (NADH, $FADH_2$), and **oxidative phosphorylation** in mitochondria uses them to pump protons across the inner membrane and drive ATP synthase. **ATP** is the energy currency that couples energy-releasing to energy-consuming reactions. Biosynthetic pathways branch from these central routes. Pathways are described in databases such as KEGG and Reactome (Section 2.6), and can be reconstructed for any sequenced genome by orthology.

For metabolism, **constraint-based modelling** avoids the need for kinetic parameters. A **genome-scale metabolic model** lists all reactions encoded in a genome, with their stoichiometry, collected in a matrix S (metabolites × reactions) (Thiele and Palsson 2010). **Flux balance analysis** (FBA) assumes a steady state, in which each metabolite is produced and consumed at equal rates, $S\mathbf{v} = 0$, where $\mathbf{v}$ is the vector of reaction fluxes. Fluxes are bounded by enzyme capacities and nutrient uptake, $\mathbf{v}_{\min} \le \mathbf{v} \le \mathbf{v}_{\max}$, and among all feasible flux distributions, linear programming finds the one that maximises an objective, usually the production of biomass (growth) (Orth et al. 2010). FBA predicted the growth of *E. coli* on different substrates and the effects of gene deletions (Edwards and Palsson 2000), and *E. coli* evolved in the laboratory towards the optimal growth rates predicted by FBA (Ibarra et al. 2002). Human metabolic models, such as Recon3D (Brunk et al. 2018) and Human-GEM from Chalmers in Gothenburg (Robinson et al. 2020), are used to study metabolic diseases and cancer metabolism, often after tailoring the model to a cell type with expression data.

11.3.7 Dynamic models of regulatory and signalling networks

Flux balance analysis describes a steady state and needs no kinetic parameters, but it cannot say how a system changes in time, or why it switches, oscillates or adapts. **Kinetic models** describe the concentration of each species with ordinary differential equations (ODEs). Under **mass action**, the rate of a reaction is proportional to the product of the concentrations of the reactants: for $A + B \rightarrow C$ with rate constant k, $d[C]/dt = k[A][B]$. For an enzyme-catalysed reaction in which the enzyme is much less abundant than the substrate, **Michaelis–Menten** kinetics gives the rate $v = V_{\max}[S]/(K_M + [S])$, which is linear at low substrate concentration and saturates at high concentration. Cooperative binding, as in gene regulation by multimeric transcription factors, is often described by a Hill function, $[X]^n/(K^n + [X]^n)$, which becomes switch-like as n increases.

The behaviour of such models depends on the structure of the network more than on the exact parameters:

- **Network motifs.** The *feed-forward loop*, in which X regulates Y and both regulate Z, is over-represented in the transcriptional networks of *E. coli* and yeast (Shen-Orr et al. 2002; Milo et al. 2002). In its coherent form, with an AND gate at Z, it responds only to a persistent signal in X and filters out brief fluctuations; in its incoherent form it can produce a pulse and speed up responses (Mangan and Alon 2003; Alon 2007).
- **Negative feedback** stabilises a level and speeds up the response, but with a delay, or several steps around the loop, it produces **oscillations**. The *repressilator*, three repressors that repress each other in a cycle, was built in *E. coli* and oscillated (Elowitz and Leibler 2000).
- **Positive feedback**, or two mutual repressors, can create **bistability**: two stable states, with the system staying in one until a large enough signal flips it, which gives memory and all-or-none cell fate decisions. A synthetic *toggle switch* demonstrated this (Gardner et al. 2000). Such small circuits have been classified by the behaviours they can produce (Tyson et al. 2003).

Biological networks are also **robust**: some properties are insensitive to large changes in parameters. In bacterial chemotaxis, precise adaptation to a constant stimulus follows from the structure of the network, not from fine-tuned rate constants (Barkai and Leibler 1997; Alon et al. 1999). Robustness arises from feedback, from **redundancy** (as in the buffering genetic interactions of Section 11.2) and from **modularity**, the organisation of the cell into semi-independent functional units (Hartwell et al. 1999; Kitano 2004). It has a cost: a system robust to common perturbations can be fragile to rare ones.

The practical difficulty is **parameter estimation**. Most rate constants are unknown and are fitted to time-course data, typically by minimising the squared difference between simulated and measured concentrations. A parameter is *structurally non-identifiable* when different values give exactly the same output for the observed species, and *practically non-identifiable* when the data are too few or too noisy to pin it down (Raue et al. 2009). Typical models are "sloppy": predictions depend on a few combinations of parameters, while most individual parameters are poorly determined (Gutenkunst et al. 2007). Useful predictions can still be made, but fitted values should not be read as measurements, and models with many parameters relative to the data should be treated with caution, as in any fitting problem (4).

ODEs describe average concentrations and are **deterministic**. When molecules are few, as for a gene present in one or two copies and its mRNA, random timing of individual reactions matters: genetically identical cells in the same environment differ in expression (Elowitz et al. 2002). **Stochastic models** are simulated with the Gillespie algorithm, which draws the time and identity of the next reaction from their exact probabilities (Gillespie 1977). Noise can also be useful, for example to make some cells in a population switch state.

Models are exchanged in the **Systems Biology Markup Language** (SBML), an XML format that describes species, reactions, rate laws and parameters, so that a model can be simulated by different software (Hucka et al. 2003). **BioModels** collects published models in SBML, many curated so that they

reproduce the results in the original paper (Malik-Sheriff et al. 2019). Compared with constraint-based models, kinetic models are much smaller (usually tens of species rather than thousands of reactions), but they describe dynamics, regulation and responses to signals that a steady-state flux analysis cannot.

See also: further explanation in Section D.11.2; related Wikipedia articles in Section E.11.2.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Interactome, Biological network, Scale-free network, Clustering coefficient, Network medicine, Systems biology, Metabolism and Flux balance analysis (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Networks, Pathways, and Enrichment" (CC BY 4.0), used with permission.

11.4 Prediction methods

Computational methods address two related questions: **which** proteins interact, and **how**, that is, the structure of the complex. Until recently these were separate fields, using genomic and evolutionary signals for the first question and docking for the second. Deep learning has merged them: methods such as AlphaFold now predict whether two proteins interact by predicting the structure of their complex.

11.4.1 Interfaces

Protein–protein interfaces were characterised from the first structures of complexes (Chothia and Janin 1975; Jones and Thornton 1996). Stable interfaces typically bury 1,200–2,000 Å² of surface, are more hydrophobic than the rest of the surface, and are closely packed. Transient interfaces are smaller and more polar. The binding energy is not spread evenly: a few **hot-spot** residues, often tryptophan, arginine and tyrosine at the centre of the interface, contribute most of it, as shown by alanine scanning (Bogan and Thorn 1998). Interface residues are more conserved than the rest of the surface, although less than the core, and this conservation is one signal used to predict interfaces.

Not all association happens through a single, well-defined interface. **Biomolecular condensates**, such as nucleoli and stress granules, are membrane-less compartments that form by phase separation, driven by many weak, *multivalent* interactions, for example between intrinsically disordered regions or between repeated domains and the short motifs they bind. Such collective assemblies are not well described by a set of binary edges in a network or by one predicted complex structure (Banani et al. 2017).

11.4.2 Genomic context

The genomic-context methods of Section 10.4, gene fusion, conserved gene neighbourhood and phylogenetic profiles, predict **functional links**, which are often, but not always, physical interactions. They are cheap, work best in prokaryotes, and are a main source of evidence in STRING (Section 11.3).

11.4.3 Coevolution between proteins

The coevolution methods used for structure prediction (Section 9.8) can also be applied to two proteins. If residues in one protein contact residues in another, mutations at the interface should be correlated across species. This requires a **paired** (concatenated) multiple sequence alignment, in which each row contains the sequences of the two proteins from the same species, and in which the two sequences actually interact. Pairing is easy for proteins encoded next to each other in bacterial operons, but hard when there are several paralogues in each genome, since it is not known which copy interacts with which.

- Correlated mutations between interacting proteins were exploited with Bayesian methods to pair interaction partners among paralogues (Burger and Nimwegen 2008). Pairing can be inferred iteratively, by choosing the pairing that maximises coevolution (Bitbol et al. 2016).
- Direct-coupling analysis of paired alignments predicted residue contacts across interfaces accurately enough to model complexes, in bacteria (Ovchinnikov et al. 2014; Hopf et al. 2014).
- Proteome-wide coevolution screens identified new interactions in bacteria (Cong et al. 2019; Green et al. 2021) and, combined with deep learning, structures of core eukaryotic complexes in yeast (Humphreys et al. 2021).

Coevolution signals are weak for interactions that are not conserved over long evolutionary distances, such as those between host and pathogen proteins and between antibodies and antigens, and for eukaryotic proteins with few homologues.

11.4.4 Docking

Protein–protein docking predicts the structure of a complex from the structures of its components. Rigid-body docking searches all relative rotations and translations of the two proteins and scores surface complementarity. The fast Fourier transform made an exhaustive search of translations efficient (Katchalski-Katzir et al. 1992), and is used by programs such as ZDOCK (Pierce et al. 2014) and ClusPro (Kozakov et al. 2017). The top-scoring poses are then refined, for example with RosettaDock, which optimises the rigid-body position and side chains together (Gray et al. 2003). HADDOCK uses experimental information, such as mutagenesis, NMR or cross-linking data, as restraints to guide docking (Dominguez et al. 2003).

Docking has been assessed in the blind **CAPRI** experiment since 2001 (Janin et al. 2003), with models scored by the fraction of native contacts and the interface and ligand RMSD, combined in the DockQ score (Basu and Wallner 2016) (Section 9.2). Rigid docking works when the unbound structures are similar to the bound ones, but fails when binding involves large conformational changes, and scoring functions struggle to pick the correct pose among many decoys.

11.4.5 Deep learning: AlphaFold for complexes

Soon after AlphaFold2 was released, it turned out that it could predict complexes when two sequences were joined with a linker, using paired alignments. Two developments followed in 2021–2022:

- **AlphaFold-Multimer** was trained specifically on complexes. Its confidence score for interfaces, **ipTM**, measures how reliable the predicted relative positions of the chains are (Evans et al. 2021).
- **FoldDock** evaluated AlphaFold2 on heterodimers, and introduced **pDockQ**, a score based on the number of interface contacts and their pLDDT, which predicts the DockQ of a model and separates interacting from non-interacting pairs (Bryant, Pozzati, and Elofsson 2022). Combining folding and docking in one step was shown to outperform docking of separately predicted structures (Pozzati et al. 2021).

Since then, confidence scores of predicted complexes have been used to **screen for interactions at proteome scale**: more than eight million pairs of yeast proteins were screened with coevolution and deep learning (Humphreys et al. 2021), tens of thousands of human pairs from interaction databases were modelled (Burke et al. 2023), and proteins of human pathogens have been screened in the same way (Humphreys et al. 2024). Large complexes that are too big to predict at once can be assembled from predicted subcomplexes, for example with Monte Carlo tree search (Bryant, Pozzati, Zhu, et al. 2022). AlphaFold3 and the open co-folding models extend this to complexes with nucleic acids and ligands (Section 9.8).

11.4.6 Evaluation and limitations

Evaluating predicted interactions is difficult, because negative examples (pairs known *not* to interact) are rarely available. That an interaction was not detected does not make a pair a true negative, unless the assay was able to detect that pair under the relevant conditions: the proteins may not have been expressed, may have been in the wrong compartment, or the interaction may be too weak or transient for the method. Random protein pairs are therefore used as a proxy, but they are unrealistically easy negatives, since they often come from different compartments or processes, and they inflate the apparent performance (Park and Marcotte 2012). Performance is also inflated when the same proteins, or close homologues of them, occur in both training and test pairs, because a method can then learn which proteins interact often rather than which pairs interact; see 4 for this kind of data leakage and how to split data to avoid it (Bernett et al. 2024).

Since only a small fraction of all possible pairs interact, even a method with a low false-positive rate gives many false positives in a proteome-wide screen, so precision at the top of the ranking matters more than overall accuracy. A high-confidence predicted interface is not proof of a biological interaction either. Co-folding methods can produce plausible, confident interfaces for proteins that never meet in the cell, for example because they are expressed in different cell types or compartments, or for paralogues of the true partner. Predicted interactions should therefore be evaluated with appropriate negatives and controls, and interpreted together with biological context such as expression and localisation. AlphaFold-based methods work well for stable, obligate complexes with coevolution between the partners, and much less well for antibody–antigen and other non-conserved interactions, for transient interactions mediated by short linear motifs in disordered regions, and for large flexible assemblies (Zhu et al. 2023). Predicted interactions remain hypotheses, but with a structural model and a confidence score they are much easier to test than earlier predictions.

See also: further explanation in Section D.11.3; related Wikipedia articles in Section E.11.3.

> **i** Sources
>
> Parts of this section are adapted from the Wikipedia articles Protein–protein interaction prediction, Protein–protein interaction, Docking (molecular), Biomolecular condensate and AlphaFold (CC BY-SA 4.0, retrieved 4 October 2026).

11.5 Key concepts

A checklist for revision.

- **Kinds of interaction:** physical binding, co-complex membership, genetic interaction, regulatory influence and functional association are not interchangeable

- **Yeast two-hybrid:** principle; binary interactions (enriched for direct contacts, but not proof of them); proteome-scale screens in yeast and human; HuRI; precision vs. sensitivity; biases (nucleus, membrane proteins, auto-activation); split-ubiquitin and PCA
- **AP-MS:** TAP tags; co-complex membership (vs. binary Y2H); spoke vs. matrix models; contaminants; BioPlex; co-fractionation, proximity labelling and cross-linking MS (residue-level distance restraints); hu.MAP
- **Binding and validation:** $K_d = k_{\text{off}}/k_{\text{on}}$ and fraction bound; SPR, BLI and ITC; FRET ($E = 1/(1 + (r/R_0)^6)$) and BRET; proximity labelling windows (BioID, TurboID, APEX); orthogonal validation with different failure modes; interface mutants and rescue
- **Quality of interaction data:** low overlap and what it means; literature and study bias; context dependence (no single fixed interactome); affinity (K_d), kinetics and stoichiometry hidden by binary edges
- **Genetic interactions:** synthetic lethality; the multiplicative null model and $\varepsilon_{AB} = w_{AB} - w_A w_B$; aggravating vs. alleviating interactions; genetic interaction profiles; the yeast genetic interaction map
- **Gene regulatory networks:** directed, signed edges; correlation and mutual information (ARACNE); GENIE3; inference from perturbations; Perturb-seq; DREAM and the wisdom of crowds
- **Network concepts:** degree distribution; random vs. scale-free networks and preferential attachment; small worlds and clustering coefficient ($C_i = 2e_i/[k_i(k_i - 1)]$); betweenness centrality; modularity; modules and community detection (Girvan–Newman, MCL); motifs
- **Hubs:** centrality–lethality; party and date hubs; degree correlations; critiques (sampling, study bias, alternative heavy-tailed distributions, "scale-free networks are rare"); correlation, not causation
- **Functional association networks:** types of evidence; gold standards; likelihood ratios and naive Bayesian integration; FunCoup and STRING; combining channel scores ($1 - \prod_i (1 - s_i)$); links mean "work together"
- **Pathway and network enrichment:** gene sets from GO, KEGG and Reactome; ORA and GSEA (Section 7.5); dependence in the GO graph (topGO elim/weight, parent–child); multiple testing of overlapping sets; annotation bias and circularity; topology-based pathway analysis (SPIA); network enrichment and crosstalk (NEA, BinoX, ANUBIX, PathBIX) with degree-aware null models; MaxLink
- **Network medicine:** the human disease network; disease modules; random walks with restart and network propagation for gene prioritisation, and how to evaluate them; active subnetworks
- **Systems biology:** metabolism in outline (glycolysis, citric acid cycle, oxidative phosphorylation, ATP); kinetic vs. constraint-based models; genome-scale metabolic models; flux balance analysis ($S\mathbf{v} = 0$, bounds, linear programming, biomass objective); Recon3D and Human-GEM
- **Dynamic models:** ODE models with mass action, Michaelis–Menten ($v = V_{\max}[S]/(K_M + [S])$) and Hill kinetics; coherent and incoherent feed-forward loops (persistence detection, pulses); negative feedback and oscillations (the repressilator); positive feedback and bistability (the toggle switch); robustness (chemotaxis adaptation), redundancy and modularity; parameter fitting, structural vs. practical identifiability and sloppiness; deterministic vs. stochastic models and the Gillespie algorithm; SBML and BioModels; kinetic vs. constraint-based models
- **Interfaces:** size, composition and hot spots; condensates and multivalent association
- **Coevolution between proteins:** paired alignments and the pairing problem; DCA across interfaces; proteome-wide screens
- **Docking:** FFT-based rigid-body search; refinement; restraint-driven docking (HADDOCK); CAPRI and DockQ; conformational change
- **Deep learning:** AlphaFold-Multimer and ipTM; FoldDock and pDockQ; proteome-scale screening; assembling large complexes; limitations (antibodies, transient and motif-mediated interactions); a confident interface is not proof of interaction; evaluation without true negatives (undetected ≠

negative; random pairs are easy negatives; shared proteins and homologues between training and test)

11.6 Recommended reading

These papers are recommended for reading in full:

- Mering et al. (2002) — comparative assessment of large-scale interaction data; why different methods disagree.
- Luck et al. (2020) — HuRI, the reference map of the human binary interactome, and how its quality was assessed.
- Barabási and Oltvai (2004) — network biology; the concepts of degree, hubs, modules and motifs.
- Orth et al. (2010) — what flux balance analysis is, in a short primer.
- Bryant, Pozzati, and Elofsson (2022) — predicting protein–protein interactions and complex structures with AlphaFold2 (FoldDock, pDockQ).
- Burke et al. (2023) — towards a structurally resolved human protein interaction network.

Titles, abstracts, links and the full texts are in Section A.11; a longer list of further papers is in Section B.11.

11.7 Background reading (open access)

- **Protein–protein interaction (Wikipedia)** — Protein–protein interactions (PPIs) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect.
  https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction
- **Two-hybrid screening (Wikipedia)** — Two-hybrid screening is a molecular biology technique used to discover protein–protein interactions (PPIs) and protein–DNA interactions by testing for physical interactions between two proteins or a single protein and a DNA molecule, respectively.
  https://en.wikipedia.org/wiki/Two-hybrid_screening
- **Interactome (Wikipedia)** — In molecular biology, an interactome is the whole set of molecular interactions in a particular cell. The term specifically refers to physical interactions among molecules but can also describe sets of indirect interactions among genes.
  https://en.wikipedia.org/wiki/Interactome
- **Biological network (Wikipedia)** — A biological network is a method of representing systems as complex sets of binary interactions or relations between various biological entities. In general, networks or graphs are used to capture relationships between entities or objects.
  https://en.wikipedia.org/wiki/Biological_network
- **Clustering coefficient (Wikipedia)** — In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.
  https://en.wikipedia.org/wiki/Clustering_coefficient
- **Scale-free network (Wikipedia)** — A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction P(k) of nodes in the network having k connections to other nodes goes for large values of k as
  https://en.wikipedia.org/wiki/Scale-free_network
- **Gene regulatory network (Wikipedia)** — A gene regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins which, in turn, determine the function of the

cell.
https://en.wikipedia.org/wiki/Gene_regulatory_network

- **Network medicine (Wikipedia)** — Network medicine is the application of network science towards identifying, preventing, and treating diseases. This field focuses on using network topology and network dynamics towards identifying diseases and developing medical drugs.
https://en.wikipedia.org/wiki/Network_medicine
- **Systems biology (Wikipedia)** — Systems biology is the computational and mathematical analysis and modeling of complex biological systems. It is a biology-based interdisciplinary field of study that focuses on complex interactions within biological systems, using a holistic approach to biological research.
https://en.wikipedia.org/wiki/Systems_biology
- **Flux balance analysis (Wikipedia)** — In biochemistry and systems biology, flux balance analysis (FBA) is a mathematical method for simulating the metabolism of cells or entire unicellular organisms, such as E. coli or yeast, using genome-scale reconstructions of metabolic networks.
https://en.wikipedia.org/wiki/Flux_balance_analysis

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

11.8 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

11.9 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Molecular Interactions: Experiments, Structures, and Networks" and "Networks, Pathways, and Enrichment" of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

11.9.1 Define the biological relation

An interaction database contains an edge between proteins A and B. What can and cannot be inferred from this edge before the underlying publication is consulted?

Extensions, solution outline and common errors

Extensions:

1. Distinguish direct contact, co-complex membership, spatial proximity, genetic interaction and functional association.
2. Which metadata are needed to interpret the edge?
3. How would the conclusion change if A and B were specific isoforms or modified forms?

Solution outline:

1. The edge alone establishes only that the database records some relation between two

identifiers.
2. Recover the molecular forms, organism, cell type, compartment, perturbation, time, detection method, confidence and publication. Curated databases such as IntAct record these in the PSI-MI standard (Section 2.7).
3. Direct binding requires evidence that excludes an unobserved bridging protein; co-complex evidence (AP-MS) allows one. Proximity records presence within an assay-specific distance and time window. A genetic interaction is a deviation of a double-mutant phenotype from a null model, and functional association may come from co-expression, pathway membership, genomic neighbourhood or the literature.
4. Several databases may repeat the same source, and are then not independent evidence.

Common errors:

- Treating every edge as direct physical binding.
- Interpreting colocalisation as molecular contact.
- Ignoring isoforms, modifications, compartment and the experimental host.
- Treating the absence of an edge as evidence of non-interaction.

Connections: Section 11.2 and Figure 11.1; interaction databases in Section 2.7; functional association networks in Section 11.3.

11.9.2 Evaluate a yeast two-hybrid result

Design a yeast two-hybrid experiment for proteins A and B, and state the strongest conclusion supported by a positive reporter signal.

Extensions, solution outline and common errors

Extensions:

1. Explain bait auto-activation and the effects of fusion orientation.
2. Identify important causes of false negatives.
3. Propose an orthogonal validation strategy.

Solution outline:

1. Fuse A to the DNA-binding domain and B to the activation domain of a transcription factor; association of the fusions reconstitutes reporter transcription, as in the original Gal4 system (Fields and Song 1989).
2. Controls: empty vectors, unrelated proteins, auto-activation tests, expression measurements, sequencing of constructs, several reporters, reciprocal orientations and independent transformants.
3. A positive result supports association of the two fusion proteins in the yeast nucleus. It does not establish direct binding in the native context: a yeast protein may bridge them, and the fusion geometry, location or expression level may create the signal.
4. Negative results can come from misfolding, toxicity, missing modifications, membrane topology or a hidden binding site.
5. Validate with a method based on a different physical principle, for example co-immunoprecipitation of the endogenous proteins and, if direct binding is the claim, binding of the purified proteins.

Common errors:

- Describing Y2H as an in vitro assay.
- Omitting auto-activation and expression controls.
- Equating a negative result with biological absence.
- Calling a closely related complementation assay an orthogonal validation.

Connections: Section 11.2; sensitivity, precision and missing data in 4.

11.9.3 Infer co-complex membership from AP-MS

Protein B is identified by mass spectrometry after affinity purification of protein A. Formulate a statistical and experimental analysis that distinguishes bait-specific enrichment from background.

Extensions, solution outline and common errors

Extensions:

1. Define suitable biological and process controls.
2. What does a SAINT score represent?
3. Does reciprocal co-immunoprecipitation prove direct contact?

Solution outline:

1. The observation makes B a candidate co-complex member.
2. Compare normalised prey abundance with matched controls (untagged cells, tag only, an irrelevant bait or a knockout) across independent biological preparations, and record bait recovery, tag position, buffer and batch.
3. SAINT estimates the probability of a true interaction from the abundance distributions in bait and control purifications (Choi et al. 2010); the CRAPome lists recurrent contaminants but cannot replace matched controls (Mellacheruvu et al. 2013).
4. Lysis may destroy weak associations or let proteins from different compartments meet.
5. Reciprocal capture strengthens the evidence for co-complex membership, but a third protein may still bridge A and B.

Common errors:

- Accepting one peptide or one purification as sufficient evidence.
- Treating repeated injections as biological replicates.
- Using controls made with different tags, buffers or cells without qualification.
- Expanding every bait–prey result into all prey–prey contacts (the matrix model).

Connections: Section 11.2; interaction proteomics and scoring in Section 8.5; peptide identification in Section 8.3.

11.9.4 Interpret proximity labelling

Compare TurboID, APEX and AP-MS for identifying the proteins associated with a membrane receptor after stimulation.

Extensions, solution outline and common errors

Extensions:

1. How do the labelling window and the orientation of the enzyme affect the result?
2. Design controls that separate receptor-specific proximity from compartmental background.
3. Propose validation for a low-abundance candidate.

Solution outline:

1. TurboID (a biotin ligase) and APEX (a peroxidase) label proteins covalently in living cells before lysis, so weak, transient, insoluble or membrane-proximal neighbours can be recovered after stringent extraction (Branon et al. 2018; Rhee et al. 2013). AP-MS recovers only complexes that survive lysis.
2. The result is a proximity proteome integrated over the labelling window (minutes for TurboID and APEX, hours for BioID), not a list of binders (Roux et al. 2012). Labelling depends on accessible residues, enzyme orientation, residence time and abundance.
3. Controls: the enzyme alone targeted to the same compartment, omission of substrate or peroxide, unstimulated cells, and tests that the fusion keeps the receptor's localisation and function.
4. Validate candidates with an orthogonal live-cell or endogenous capture method and, for direct binding, with purified proteins or interface mutations.

Common errors:

- Assuming one universal labelling radius.
- Calling every enriched protein a direct interactor.
- Ignoring abundant bystanders in the same compartment and global changes in abundance.

Connections: Section 11.2; proximity labelling and quantitative proteomics in Section 8.5; localisation in 3.

11.9.5 Distinguish proximity from binding with FRET

A FRET signal increases after ligand stimulation. Give the relevant relationship, and state the biological interpretations compatible with the observation.

Extensions, solution outline and common errors

Extensions:

1. What are the roles of donor–acceptor orientation and the Förster distance?
2. Compare FRET with BRET.
3. Design controls for crowding, expression level and conformational change.

Solution outline:

1. The efficiency is $E = 1/[1 + (r/R_0)^6]$, where r is the donor–acceptor distance and R_0 the distance at which $E = 0.5$ (Stryer and Haugland 1967).
2. The increase shows a change in the proximity or orientation of the fluorophores, or in the populations of states. It does not distinguish direct binding from crowding, recruitment to a larger complex, or a conformational change within one protein.

3. BRET uses a luciferase as donor, avoiding excitation light and autofluorescence.
4. Controls: donor-only and acceptor-only samples, spectral correction or lifetime measurements, titration of expression and acceptor saturation, non-binding proteins in the same compartment, tag swaps, and interface mutants that remain folded.

Common errors:

- Treating FRET as an exact distance measurement.
- Concluding that the absence of FRET excludes binding.
- Using a single overexpression level without testing for saturation.

Connections: Section 11.2; conformational ensembles in Section 9.11.

11.9.6 Use cross-linking MS as a structural restraint

Explain how cross-linking mass spectrometry constrains the structure of a protein complex, and why the cross-linker is not an exact molecular ruler.

Extensions, solution outline and common errors

Extensions:

1. Why is the identification of cross-linked peptide pairs harder than ordinary peptide identification?
2. How can cross-links be combined with docking or integrative modelling?

Solution outline:

1. A chemical cross-linker covalently connects reactive side chains that are close in the sample; after digestion and tandem MS, linked peptide pairs are identified and mapped to residues (O'Reilly and Rappsilber 2018).
2. With n candidate peptides there are of the order of n^2 possible pairs, so the search space is much larger and false discoveries must be controlled carefully.
3. A cross-link is compatible with a range of distances (up to about 30 Å between Cα atoms for common reagents), depending on spacer length, side chains, flexibility and conformational state.
4. Consistent cross-links can filter docking poses or act as restraints in docking (HADDOCK, Section 11.4) and integrative modelling (Section 9.4), but alternative stoichiometries and conformations must be tested.
5. A missing cross-link is weak negative evidence, since reactive residues and detectable peptides are unevenly distributed.

Common errors:

- Using the nominal spacer length as an exact distance cutoff.
- Ignoring aggregation and mixtures of states.
- Choosing one structural model without testing alternatives.

Connections: Section 11.2 and Section 11.4; XL-MS in Section 8.5; integrative structures in Section 9.4.

11.9.7 Evaluate a predicted complex

An AlphaFold-class model predicts a high-confidence interface between A and B. Design an evaluation that separates structural plausibility from evidence that the proteins associate in the cell.

Extensions, solution outline and common errors

Extensions:

1. Which confidence and physical-quality measures should be inspected?
2. Design biologically realistic negative controls.
3. How would you benchmark a learned interaction predictor without homology leakage?

Solution outline:

1. The model is a structural hypothesis conditional on having submitted both partners. Inspect ipTM, inter-chain PAE, interface pLDDT or pDockQ, clashes, buried chemistry, conservation, repeated seeds and alternative stoichiometries.
2. Compare with non-partners matched for localisation and size, competing paralogues, and known positive and tested negative pairs. A co-folding score is not a dissociation constant.
3. Coevolution between the chains supports a contact only if orthologues and paralogues are paired correctly and the alignment is deep enough.
4. Benchmarks should separate homologous families, complexes or time periods between training and test sets; random splits of pairs leak information through shared proteins. Unreported pairs are unknown, not verified negatives (Park and Marcotte 2012; Bernett et al. 2024).

Common errors:

- Reading confidence as interaction probability or affinity.
- Evaluating only on known positive complexes.
- Ignoring incompatible compartments and competing partners.

Connections: Section 11.4; AlphaFold in Section 9.8; leakage and calibration in 4.

11.9.8 Analyse physical and genetic networks without mixing their meaning

A protein has the highest degree in an integrated interaction network, and a pair of genes shows a strongly negative genetic interaction. What do these observations imply, and how should they be validated?

Extensions, solution outline and common errors

Extensions:

1. Define degree, betweenness, modularity and the multiplicative genetic-interaction score.
2. Explain study bias, spoke and matrix expansion, and unequal detectability.
3. Design an analysis comparing the network before and after stimulation.

Solution outline:

1. Degree $k_i = \sum_j A_{ij}$ is observed connectivity in one graph; betweenness counts shortest paths

through a node; modularity compares edges within modules with a degree-preserving null. These describe the graph, not binding, causal control or essentiality.
2. Abundant, easily studied and well-known proteins are tested more often; matrix expansion of AP-MS data adds prey–prey edges that need not be contacts. Both inflate degree.
3. For fitness values, $\varepsilon_{AB} = w_{AB} - w_A w_B$; a negative value means a double-mutant phenotype more severe than the multiplicative expectation, often for genes in parallel pathways, not protein binding (Costanzo et al. 2016).
4. Comparisons between conditions should model quantitative evidence and account for bait and prey abundance, batch and missing values, rather than subtract thresholded edge lists.
5. Validate genetic interactions by reconstructing the double mutant; physical mechanisms need independent molecular measurements.

Common errors:

- Mixing direct, co-complex, proximity, genetic and functional edges without labels.
- Treating centrality as indispensability (the centrality–lethality debate).
- Interpreting a network module as a biochemical complex without validation.

Connections: Section 11.3 and Section 11.2; graph null models in 4.

11.9.9 When is a network a biological model?

Compare the meaning of a co-expression network, a protein interaction network, a pathway graph and a metabolic network. When does each support mechanistic inference, and how does pathway enrichment differ from a model?

Extensions, solution outline and common errors

Extensions:

- What does an edge mean in each network?
- Why can hub genes be artefacts?
- How does network enrichment analysis differ from over-representation analysis, and why does it need a degree-preserving null model?
- Explain flux balance analysis.

Solution outline:

1. For each network, define the nodes, the meaning of the edges, their direction, confidence, context and provenance: statistical dependence, physical evidence, curated reactions, or stoichiometric participation.
2. Well-studied genes receive more annotations and edges, which creates hubs and biases enrichment.
3. Over-representation analysis tests the overlap of a gene list with a gene set against a background; network enrichment tests whether the list has more links to the set than expected in randomised networks with the same degrees, which prevents hubs and large pathways from appearing significant by chance.
4. Enrichment summarises a list against prior knowledge; a model predicts states or fluxes. Flux balance analysis maximises an objective $\mathbf{c}^T\mathbf{v}$ subject to $S\mathbf{v} = 0$ and flux bounds; the objective and the reconstruction are assumptions.
5. Validate network conclusions with degree-preserving null models, sensitivity to the database and release, and experiments appropriate to the claim.

Common errors:

- Calling a correlation a physical interaction or regulation.
- Assuming that pathway boundaries are unique or complete.
- Saying that FBA measures fluxes or proves that cells maximise growth.
- Testing against a random graph that does not preserve degrees.

Connections: Section 11.3 (network concepts, pathway and network enrichment, metabolic modelling); over-representation analysis and GSEA in Section 7.5; graph neural networks in Section 12.4.

11.9.10 Building a kinetic model of a feedback loop

A transcription factor X represses its own gene. A time course after induction shows that X rises, overshoots and then settles at a steady level. Write a simple ODE model, explain which feature produces the overshoot, and discuss how you would fit and test it.

Extensions, solution outline and common errors

Extensions:

- How would the behaviour change if X activated its own gene instead?
- Which parameters do you expect to be identifiable from measurements of X alone?

Solution outline:

1. A minimal model: $dm/dt = \alpha/(1 + (X/K)^n) - \delta_m m$ for the mRNA and $dX/dt = \beta m - \delta_X X$ for the protein, with a Hill function for repression.
2. Negative autoregulation speeds up the approach to steady state and reduces its variability. The overshoot arises from the delay between transcription and the accumulation of protein: the repression acts on mRNA that is already made. A steep Hill function (large n) and long delays push the system towards damped or sustained oscillations.
3. Fit the parameters by least squares to the time course, ideally with several perturbations (induction strengths, mutants). Examine identifiability with profile likelihoods: with only protein measurements, the synthesis rates and K may be confounded, so only combinations of them are determined.
4. Test the model on a prediction it was not fitted to, such as the response of a mutant with a weakened binding site.
5. Positive autoregulation with a steep Hill function can instead give bistability: two stable levels and memory of past induction.
6. If X is present in few molecules, a stochastic simulation (Gillespie) shows cell-to-cell variation that the ODE averages away.

Common errors:

- Interpreting fitted parameters as measured rate constants.
- Concluding from a good fit that the model structure is correct.
- Forgetting that the deterministic model describes a population average.

Connections: Section 11.3 (dynamic models, systems biology); gene regulatory networks in Section 11.2; model fitting and overfitting in 4.

12 Machine learning

> **From the exam description**
>
> Fundamental knowledge about machine learning. How machine learning is used in bioinformatics. Recent progress in machine learning methodologies.

12.1 Overview

Machine learning is now part of almost every area of bioinformatics, and recent progress in deep learning has changed several of them completely. This chapter cannot include everything. It covers the minimum that every bioinformatician should understand, which corresponds roughly to the course *Machine learning for bioinformatics* (KB8029), and points to the applications described in the other chapters:

- the basics: kinds of learning, linear models and regularisation, generalisation and overfitting, the pitfalls of biological data, and how to evaluate predictors (4);
- the history: the perceptron, neural networks and backpropagation, hidden Markov models, support vector machines and random forests (Section 12.3);
- deep learning: convolutional, recurrent, transformer and graph networks, language models for proteins and DNA, interpretability and reporting standards (Section 12.4);
- generative models: autoregressive models, autoencoders, GANs and diffusion (Section 12.5);
- foundation models and how to evaluate them (Section 12.6);
- AI in research practice: what large language models are, what they have achieved in biology, and how to use AI tools responsibly (Section 12.7);
- integrating data: multi-omics integration, latent factors and networks, and what integration can and cannot show about causes (Section 12.8).

Applications are found throughout the book: profile HMMs and protein language models in Chapter 1, gene finding in Chapter 5, single-cell analysis in Chapter 7, de novo peptide sequencing in Chapter 8, structure prediction and design in Chapter 9, function and variant effect prediction in Chapter 10, and image analysis and virtual cell models in Chapter 13.

12.2 Basics

Machine learning (ML) is a set of methods that learn patterns from data instead of having them programmed explicitly. Bioinformatics has used ML since the early 1980s, and it is now part of almost every area covered in this book (Tarca et al. 2007; Greener et al. 2021; Sapoval et al. 2022). This section introduces the basic concepts, and the pitfalls that are specific to biological data. It is not possible to cover the whole field here; the aim is the minimum that a bioinformatician must understand to use and evaluate ML methods.

Terminology: AI, ML, deep learning, foundation models

The terms overlap but are not synonyms.

- **Artificial intelligence (AI)** is the broad goal of machines performing tasks that are thought to require intelligence. It includes rule-based systems that do not learn from data.
- **Machine learning (ML)** is the part of AI in which a model learns its parameters from data. It includes linear and logistic regression, HMMs, SVMs and random forests.
- **Deep learning** is the part of ML that uses neural networks with many layers, which learn their own features from raw input (Section 12.4).
- **Generative AI** refers to models that generate new data, such as text, images, sequences or structures (Section 12.5). Not all generative models are deep.
- A **foundation model** is a model, usually deep and pre-trained with self-supervision on broad data, that is used for many downstream tasks (Section 12.6). The term describes how a model is used, not its size. **Large language models** (LLMs) are foundation models for text.

12.2.1 Kinds of learning

- In **supervised learning**, a model is trained on labelled examples, pairs of inputs $\mathbf{x}$ and outputs y, to predict y for new inputs. If y is a category (signal peptide or not; cell type), the task is **classification**; if it is a number (binding affinity; expression level), it is **regression**.
- In **unsupervised learning**, there are no labels. The model finds structure in the data, for example clusters or low-dimensional representations (PCA, t-SNE, UMAP; Section 7.4).
- In **self-supervised learning**, the model creates its own labels from the data, for example by hiding some residues of a protein sequence and predicting them from the rest. This makes it possible to train on hundreds of millions of unlabelled sequences, and it is the basis of language and foundation models (Section 12.6).

In all cases, a model is a function $f(\mathbf{x}; \theta)$ with parameters θ (weights), and **training** means choosing θ to minimise a **loss function** on the training data, such as the squared error for regression or the cross-entropy for classification. For most models this is done with **gradient descent**: the parameters are moved a small step against the gradient of the loss, $\theta \leftarrow \theta - \eta \nabla L$, where η is the *learning rate*, usually computed on small random batches of data (*stochastic* gradient descent) and improved with momentum-based optimisers such as Adam.

12.2.2 Linear models and regularisation

The simplest supervised models are linear, and they should be the first baseline for most problems. **Linear regression** models the outcome as a weighted sum of the features, $y = \beta_0 + \mathbf{x}^T\beta + \epsilon$, and ordinary least squares chooses the weights that minimise the squared error. **Logistic regression** models a binary outcome through its log odds,

$$\log \frac{p}{1-p} = \beta_0 + \mathbf{x}^T\beta, \qquad p = P(y = 1 \mid \mathbf{x}),$$

and is fitted by maximum likelihood, which is the same as minimising the cross-entropy loss. A coefficient β_j describes how the prediction changes with x_j when the other features in the model are held fixed; in logistic regression, e^{β_j} is an odds ratio. These are conditional associations, not causal effects, and they change if correlated features are added or removed. Linear and logistic regression are special

cases of *generalised linear models*, which also include the negative binomial models used for read counts (Section 7.4).

With many more features than examples, which is the normal situation in omics (20,000 genes and perhaps 100 patients), an unpenalised linear model fits the training data perfectly and generalises poorly. **Regularisation** adds a penalty on the size of the weights to the loss, $L(\beta) + \lambda P(\beta)$, and the strength λ is a hyperparameter:

- **L2** (*ridge*) regression uses $P = \sum_j \beta_j^2$. It shrinks all weights smoothly towards zero but rarely sets any to exactly zero, and it stabilises the fit when features are correlated by spreading the weight over them (Hoerl and Kennard 1970).
- **L1** (the *lasso*) uses $P = \sum_j |\beta_j|$. Because the penalty has corners on the axes, many weights become exactly zero, so the lasso also *selects* features and gives a sparse model (Tibshirani 1996).
- The **elastic net** combines the two penalties. It selects features like the lasso, but tends to keep or drop groups of correlated features together (Zou and Hastie 2005).

A common mistake is to reverse L1 and L2: it is L1 that gives sparse models. The penalties act on the size of the weights, so features must be put on comparable scales, for example standardised, before fitting, and λ must be chosen inside the training data, by (nested) cross-validation (below). A sparse "signature" of genes selected by the lasso is a predictive model, not a list of causal genes. When genes are correlated, as genes in the same pathway usually are, the lasso picks one of them more or less arbitrarily, and a different sample of patients can give a different signature with the same performance. The stability of the selection should therefore be checked by resampling.

12.2.3 Generalisation and overfitting

The goal of a model is not to fit the training data but to make correct predictions on new data, to **generalise**. A flexible model with many parameters can memorise its training set, including its noise, and then perform poorly on new data. This is **overfitting**. A model that is too simple cannot capture the pattern at all (*underfitting*). The balance is described by the *bias–variance trade-off* (Lever et al. 2016; Hastie et al. 2009).

Overfitting is controlled by:

- **Separate data sets.** Data are split into a **training set**, used to fit the parameters, a **validation set**, used to choose hyperparameters and compare models, and a **test set**, used once at the end to estimate performance. If the test set is used for any modelling decision (choosing features, hyperparameters, architectures or which of several models to report), it has in effect become a validation set and its estimate is optimistic. A new, untouched test set is then needed.
- **Cross-validation.** With little data, the data are divided into k parts; each part is used once for evaluation while the model is trained on the rest, and the results are averaged. If hyperparameters are also chosen by cross-validation, the error of the best setting is optimistic, because it was selected on the same data. **Nested cross-validation** avoids this: an inner loop, run only on the training part of each outer fold, selects the hyperparameters, and the outer loop estimates the performance of the whole procedure, selection included. This matters most for small data sets, where the optimism of selection can be large (Varma and Simon 2006; Hastie et al. 2009).
- **Regularisation.** Penalising large weights (L2 or *weight decay*; L1, which also drives weights to zero), stopping training when the validation error starts to increase (*early stopping*), or randomly switching off units during training (*dropout*) (Srivastava et al. 2014).
- **More data**, or data augmentation.

The same idea appears in structural biology: the free R-factor of crystallography is a held-out test set (Section 9.4).

12.2.4 Pitfalls specific to bioinformatics

Biological data violate the assumption, made by most ML methods, that examples are independent. The most important consequence is that a **random split** into training and test sets is rarely appropriate. Related examples, such as homologous proteins, samples from the same patient, cells from the same donor or experiment, or genomic windows from the same chromosome, end up on both sides of a random split, and the measured performance can be grossly overestimated. Data should instead be split by **group** (all examples from one cluster, patient, donor or chromosome go to the same subset), by **time**, or by both, so that the test set resembles the new data the method will actually be used on.

- **Homology between training and test sets.** Proteins and genes are related by evolution. If a test protein has a close homologue in the training set, the model can predict it by remembering the homologue, and the measured performance says little about new families. Data sets must therefore be **redundancy-reduced**, and split so that no test example is too similar to any training example. Hobohm and colleagues introduced algorithms for selecting representative, non-redundant protein sets (Hobohm et al. 1992), and today sequences are clustered with tools such as CD-HIT or MMseqs2 (Li and Godzik 2006; Steinegger and Söding 2018), with all members of a cluster placed in the same subset. Simple clustering can leave close homologues in different subsets, and partitioning algorithms such as GraphPart construct splits in which no test sequence is above a chosen similarity to any training sequence, while keeping the classes balanced (Teufel et al. 2023). For structure, splits by fold or by date (training on structures released before a cutoff, testing on later ones) are stricter.
- **Data leakage.** Information about the test set can leak into training in less obvious ways: through features computed from the whole data set, through normalisation, through the same patient or experiment appearing in both sets, or through benchmark data that have been used to develop earlier versions of the method. Every step that learns from the data, including normalisation, imputation, feature selection, batch correction, oversampling of the rare class and calibration, is part of the model and must be fitted on the training part of each split only. A classic example is selecting the genes most correlated with the outcome on the whole data set and then cross-validating a classifier on those genes: the estimated error can be close to zero even when the labels are random (Ambroise and McLachlan 2002). A survey of many fields found that leakage of this kind was widespread in machine-learning-based science, and that correcting it often removed the claimed advantage of complex models (Kapoor and Narayanan 2023).
- **Class imbalance.** Most biological prediction problems are imbalanced: few residues are in binding sites, few variants are pathogenic, few protein pairs interact. A classifier that always predicts the majority class can have high accuracy and be useless.
- **Biased and noisy labels.** Labels come from databases with their own biases (Section 10.7), and the negatives are often not true negatives.
- **Distribution shift.** A model can perform well on held-out data from the same source and fail when the data change: a new species, protein family, hospital, sequencing technology or laboratory protocol. Performance should therefore be reported for the kind of data the model will be applied to, and a test set from a different source (*external validation*) is the best check.
- **Prediction is not causation.** A predictor may exploit any correlation in the data, including confounders such as batch, sequencing depth or ancestry, without capturing the underlying mechanism. A gene whose expression predicts a disease need not cause it, and a feature that is important for a model is not necessarily important biologically. Establishing causes requires perturbation experiments or methods designed for causal inference, such as Mendelian randomisation. Explaining and predicting are different goals, and a model that is best for one is often not best for the other (Shmueli 2010) (Section 12.8).

12.2.5 Evaluation

For a binary classifier, predictions are summarised in a **confusion matrix** of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Common measures are:

Table 12.1: Measures computed from the confusion matrix.

Measure	Definition	Question it answers
Accuracy	$(TP+TN)/(TP+FP+TN+FN)$	What fraction is correct?
Sensitivity (recall)	$TP/(TP+FN)$	What fraction of the positives is found?
Specificity	$TN/(TN+FP)$	What fraction of the negatives is correctly rejected?
Precision	$TP/(TP+FP)$	What fraction of the predicted positives is correct?
F1 score	$2\cdot\text{precision}\cdot\text{recall}/(\text{precision}+\text{recall})$	Harmonic mean of precision and recall

These measures are not interchangeable. The false-positive rate ($FP/(FP+TN) = 1-$ specificity) is computed over the negatives, whereas the false discovery rate ($FP/(TP+FP) = 1-$ precision) is computed over the predictions; with many negatives, a low false-positive rate can coexist with a high false discovery rate. **Accuracy** is often misleading for biological problems, because the classes are imbalanced: if 1% of residues are in binding sites, predicting "no" everywhere is 99% accurate. Sensitivity, specificity, precision and MCC, reported together, are the core measures.

The **Matthews correlation coefficient**,

$$\text{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

is the correlation between predicted and true labels. It ranges from -1 to 1, is 0 for random predictions, and is high only if the classifier does well on both classes, so it is more informative than accuracy or F1 for imbalanced data (Chicco and Jurman 2020). MCC was used early for secondary structure prediction.

Most classifiers output a score, and the measures above depend on the threshold. Two curves describe performance over all thresholds:

- The **ROC curve** plots sensitivity against the false-positive rate ($1 - $ specificity). The area under it (**AUC** or AUROC) is the probability that a random positive is scored higher than a random negative: 0.5 for random and 1 for perfect. The ROC curve does not depend on the proportion of positives, which is why it can look excellent when positives are rare while precision is poor.
- The **precision–recall curve** plots precision against recall, and the area under it (AUPRC or PR-AUC, close to the average precision) summarises it. A random classifier has a precision equal to the fraction of positives, not 0.5. For imbalanced problems, where the interesting question is how many of the top predictions are correct, it is more informative than the ROC curve, since a small false-positive *rate* can still mean many false positives (Saito and Rehmsmeier 2015). For example, with 100 true binding residues among 10,000, a false-positive rate of 1% gives about 100 false positives, so even if every true site is found, half of the predictions are wrong.

In practice a method is used at one threshold, so performance should also be reported at the **operating point** that will be used, for example the precision at a given recall or the sensitivity at 1% false-positive rate.

Ranking and calibration. AUC measures only how well a method *ranks* examples. It says nothing about whether the scores are probabilities. A model is **calibrated** if, among all examples given a score of 0.9, about 90% are positive. Many models are not: an output of 0.9 from a neural network or a random forest is a score, not a 90% probability, unless calibration has been checked. Calibration is assessed with a **reliability diagram** (calibration curve), which plots the observed fraction of positives against the predicted probability in bins, and can be corrected afterwards on held-out data, for example by Platt scaling (fitting a logistic function to the scores) or isotonic regression. Calibration also depends on the proportion of positives: a model calibrated on a balanced training set will overestimate probabilities when positives are rare in practice. Calibrated probabilities are needed when scores are combined with costs to choose a threshold, or interpreted as confidence for a single prediction. The quality of probabilities can be summarised with a *proper scoring rule*, which is optimised only by the true probabilities: the **Brier score**, $\frac{1}{n}\sum_i (p_i - y_i)^2$, the mean squared difference between the predicted probability and the outcome (0 or 1), or the log loss, which penalises confident errors heavily.

Discrimination, calibration and the choice of threshold answer three different questions: can the model rank positives above negatives, do its probabilities mean what they say, and which errors matter when it is used? The threshold is a decision, not a property of the model. It should be chosen from the costs of the two kinds of error, a missed pathogenic variant is usually worse than one more variant to check, using validation data rather than the test set.

Two kinds of uncertainty are often distinguished. **Aleatoric** uncertainty is noise in the data themselves, such as measurement error or genuinely variable outcomes, and does not decrease with more data. **Epistemic** uncertainty is the model's lack of knowledge, which is large for inputs unlike the training data (for example a protein from an unseen family) and can be reduced with more data. Ensembles of models that disagree on an input are a common indicator of epistemic uncertainty.

For regression, performance is measured with the correlation (Pearson or Spearman) or the mean squared error between predicted and true values.

Benchmark design. How the test data are chosen matters as much as the measure. In a **retrospective** benchmark, methods are tested on data that already exist, often data the developers could have seen; it is easy to run but prone to leakage and repeated tuning. Stricter designs are:

- a **family or cluster split**, in which test examples have no close homologue (or other related example) in the training set, which estimates performance on new families;
- a **temporal split**, in which the method is trained on data available before a cutoff date and tested on data that appeared later, which mimics real use;
- **external validation** on an independent data set, from another laboratory, technology or population, which tests robustness to distribution shift;
- a **prospective** or **blind** benchmark, in which predictions are made before the answers are known and are assessed by others.

Blind benchmarks are the most reliable way to compare methods, because the test data do not exist when the methods are developed. Bioinformatics has a strong tradition of such community experiments: CASP for structure (Section 9.2), CAFA for function (Section 10.7), CAGI for variants (Section 10.6), CAMI for metagenomics (Section 5.7) and DREAM for networks (Section 11.2). Every new method should also be compared with a **simple baseline**, such as nearest-neighbour transfer by BLAST, logistic regression or the most frequent class. Surprisingly often, sophisticated methods do not beat them by much.

See also: further explanation in Section D.12.1; related Wikipedia articles in Section E.12.1.

Sources

Parts of this section are adapted from the Wikipedia articles Machine learning, Overfitting, Cross-validation (statistics), Precision and recall, Phi coefficient, Probabilistic classification and Receiver operating characteristic (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Statistics and Machine Learning for Biological Data" (CC BY 4.0), used with permission.

12.3 A short history of machine learning in bioinformatics

Machine learning and bioinformatics grew up together. Many methods that are now standard were first applied to biological sequences in the 1980s and 1990s, and the history explains why some ideas keep returning.

12.3.1 The perceptron

In 1958 Frank Rosenblatt described the **perceptron**, a model of a neuron that computes a weighted sum of its inputs and outputs 1 if the sum exceeds a threshold (Rosenblatt 1958):

$$y = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The perceptron learning rule adjusts the weights after each mistake, and converges if the two classes can be separated by a hyperplane. A position weight matrix used to score a sequence motif (Section 1.6) is exactly such a weighted sum. In one of the first applications of ML in molecular biology, Gary Stormo and colleagues used the perceptron in 1982 to recognise translation initiation sites in *E. coli* (Stormo et al. 1982).

In 1969 Marvin Minsky and Seymour Papert showed that a single perceptron cannot learn functions that are not linearly separable, the classic example being exclusive or (XOR) (Minsky and Papert 1969). Their book contributed to a decline of neural network research for more than a decade. In biology, XOR-like patterns correspond to interactions between positions (*epistasis*), where the effect of a residue depends on another.

12.3.2 Neural networks and backpropagation

Networks with **hidden layers** of units with non-linear activation functions can represent any function, including XOR, but for a long time there was no efficient way to train them. **Backpropagation**, popularised by Rumelhart, Hinton and Williams in 1986 (Rumelhart et al. 1986), computes the gradient of the loss with respect to every weight by applying the chain rule backwards through the network, so that all weights can be trained by gradient descent (Krogh 2008).

Neural networks were applied to protein secondary structure almost immediately. Qian and Sejnowski in 1988 (Qian and Sejnowski 1988) and Holley and Karplus in 1989 (Holley and Karplus 1989) trained networks on windows of sequence, reaching about 64% accuracy. The decisive improvement came from the input, not the network: **PHD** fed evolutionary profiles from multiple alignments into a two-layer network and reached more than 70% (Rost and Sander 1993) (Section 9.8). This was an early example of a lesson that has been learned many times since: the **representation** of the data matters as much as

the model. Neural networks were also used for signal peptides (SignalP, 3) and many other sequence features in the 1990s, often by Danish groups around Søren Brunak and Anders Krogh.

12.3.3 Hidden Markov models

Hidden Markov models (HMMs) describe a sequence as generated by a series of hidden states, each emitting symbols with its own probabilities (Rabiner 1989; Eddy 2004a). They were developed for speech recognition, and introduced to biology by Churchill for DNA segmentation (Churchill 1989) and by Krogh, Haussler and colleagues for protein families (Krogh et al. 1994). Three algorithms, all based on dynamic programming, make HMMs practical:

- the **Viterbi algorithm** finds the most probable path of hidden states for a sequence (Viterbi 1967);
- the **forward–backward algorithm** computes the total probability of the sequence and the posterior probability of each state at each position;
- the **Baum–Welch algorithm**, a form of expectation–maximisation, trains the parameters from unlabelled sequences (Baum et al. 1970).

HMMs became one of the most successful ML methods in bioinformatics: profile HMMs for protein families (HMMER and Pfam, Section 1.6), gene finders (GeneMark, GENSCAN, Section 5.6) and membrane protein topology (TMHMM, Section 9.10) (Durbin et al. 1998). Their strengths are a clear probabilistic interpretation, a grammar of states that encodes biological knowledge, and the ability to handle sequences of variable length. Their weakness is that each position depends only on the current state, so long-range dependencies cannot be modelled.

HMMs differ from the neural networks and SVMs in this section. They are probabilistic **generative** models of sequences (Section 12.5), whose structure is usually designed by hand, and they are often regarded as statistics rather than machine learning. They belong in this history because their parameters are learned from data, by counting in labelled sequences or by Baum–Welch in unlabelled ones, and because they became some of the most widely used learned models in bioinformatics.

12.3.4 Support vector machines, trees and ensembles

In the 1990s and 2000s, other methods became popular for classification with hand-crafted features:

- **Support vector machines** (SVMs) find the hyperplane that separates two classes with the largest *margin* (Cortes and Vapnik 1995; Noble 2006). Through the **kernel trick**, they implicitly map data to high-dimensional spaces in which non-linear problems become linearly separable, and kernels can be defined directly on sequences or structures. SVMs were used for remote homology detection, function prediction and many other problems, and are the basis of the semi-supervised rescoring in proteomics (Section 8.4).
- **Decision trees** split the data by asking a series of questions about the features. A single tree overfits easily, but **random forests** average many trees, each trained on a bootstrap sample of the data and a random subset of the features (Breiman 2001). **Gradient boosting** builds trees sequentially, each correcting the errors of the previous ones (Friedman 2001); XGBoost is a widely used implementation (Chen and Guestrin 2016).

For small, tabular data sets with well-chosen features, random forests and gradient boosting are still often as good as deep learning, are faster to train and are easier to interpret. They should be among the first baselines tried for a new problem.

See also: further explanation in Section D.12.2; related Wikipedia articles in Section E.12.2.

Sources

Parts of this section are adapted from the Wikipedia articles Perceptron, Backpropagation, Hidden Markov model, Support vector machine and Random forest (CC BY-SA 4.0, retrieved 4 October 2026).

12.4 Deep learning

Deep learning uses neural networks with many layers that learn their own features from raw input, rather than relying on features designed by hand (LeCun et al. 2015). It became dominant after 2012, when a deep convolutional network won the ImageNet image recognition competition by a large margin (Krizhevsky et al. 2017), made possible by large data sets, graphics processors (GPUs) and better training methods. In biology it has driven some of the largest advances of the last decade, most notably AlphaFold (Ching et al. 2018; Eraslan et al. 2019; Sapoval et al. 2022).

12.4.1 Training deep networks

Deep networks are trained with backpropagation and stochastic gradient descent, like the networks of the 1980s, but several innovations made depth possible. Activation functions such as the rectified linear unit (ReLU, $\max(0, x)$) avoid the *vanishing gradients* of deep networks with sigmoid units. Careful initialisation, normalisation layers and adaptive optimisers such as Adam stabilise training. **Residual connections**, which add the input of a block to its output, allow networks with hundreds of layers to be trained (He et al. 2016). Overfitting is controlled with dropout, weight decay, early stopping and data augmentation (4).

The reason is the chain rule. The gradient for an early layer is a product of one factor per later layer, and with saturating activations such as the sigmoid or tanh most of these factors are small, so the product vanishes. The gradient of ReLU is 1 for positive inputs, and a residual connection, $\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + F(\mathbf{h}_\ell)$, gives the gradient a direct path past each block (Figure 12.1).

12.4.2 Architectures

Different architectures build different assumptions about the data into the network.

- **Convolutional neural networks** (CNNs) apply the same small filters at every position of an image or sequence (Lecun et al. 1998). Weight sharing means far fewer parameters than a fully connected network, and the filters learn local patterns, such as edges in images or motifs in DNA. Stacking layers with pooling lets later layers see larger regions. In genomics, CNNs learned the binding specificities of DNA- and RNA-binding proteins (DeepBind) (Alipanahi et al. 2015), predicted chromatin features and the effects of non-coding variants from sequence (DeepSEA, Basset) (Zhou and Troyanskaya 2015; Kelley et al. 2016), and, applied to coevolution matrices as images, predicted protein contact maps (Section 9.8).
- **Recurrent neural networks** (RNNs) process a sequence one element at a time, carrying a hidden state that summarises what has been seen. The **long short-term memory** (LSTM) unit uses gates to keep information over long distances (Hochreiter and Schmidhuber 1997). Bidirectional LSTMs read sequences in both directions, and were used in DeepLoc and early protein language models. RNNs are slow to train, because they must process positions in order.

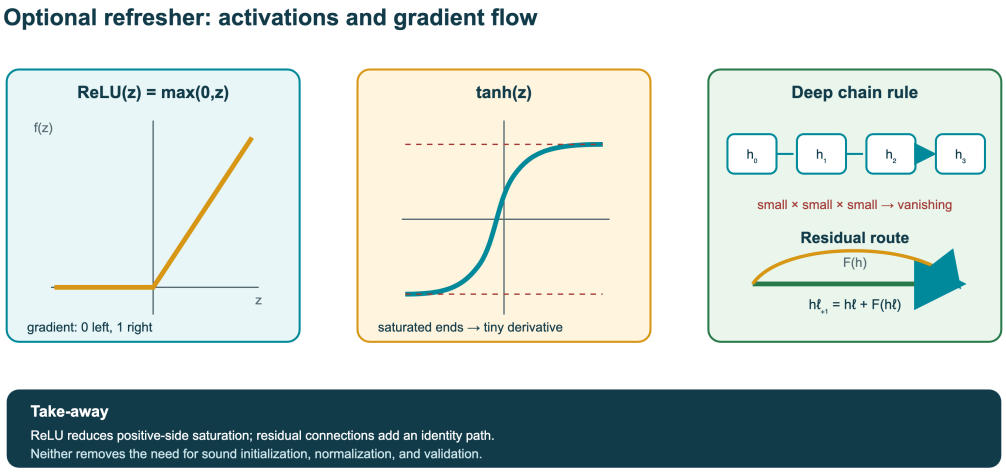


Figure 12.1: Why deep networks became trainable. Sigmoid-like activations such as tanh saturate, so their derivatives are close to zero for large inputs, and the product of many small factors in the chain rule makes the gradient vanish. ReLU has a gradient of 1 for positive inputs, and a residual connection adds an identity path around each block. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

- **Attention and transformers.** *Attention* lets each position look directly at every other position, with weights computed from the content of the positions. The **transformer** architecture is built from layers of self-attention and feed-forward networks, with residual connections and normalisation (Vaswani et al. 2017). Since every pair of positions can exchange information directly in every layer, transformers can model dependencies between distant positions, such as residues that are far apart in sequence but in contact in the structure. How well they do so in practice depends on the context length (the cost of attention grows with the square of the sequence length), the training data and the architecture. Transformers can also be trained in parallel on very large data sets. Transformers are the basis of large language models, protein language models and AlphaFold.
- **Graph neural networks** (GNNs) operate on graphs, passing *messages* between connected nodes. They are natural for molecules, protein structures (residues connected by spatial proximity) and interaction networks, and are used, for example, in structure-based function prediction and in ProteinMPNN (Section 10.4, Section 9.12). Architectures that respect the symmetries of three-dimensional space, such as AlphaFold's invariant point attention, are related.
- **Generative models**, including autoencoders, GANs and diffusion models, learn to generate new data and are described in Section 12.5.

In self-attention, each token (a residue, a base or an image patch) is represented by a vector, and three learned linear maps turn the matrix of token vectors X into *queries* Q, *keys* K and *values* V. The output is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where d_k is the dimension of the keys. Each row of the softmax gives the weights with which one token collects information from all the others (Figure 12.2). *Multi-head* attention runs several such operations in parallel with different projections. Attention by itself ignores the order of the tokens, so information about their positions must be added to the input. The $L \times L$ matrix of weights for a sequence of length L

is the source of the quadratic cost mentioned above, which matters for whole chromosomes, long reads and large sets of cells.

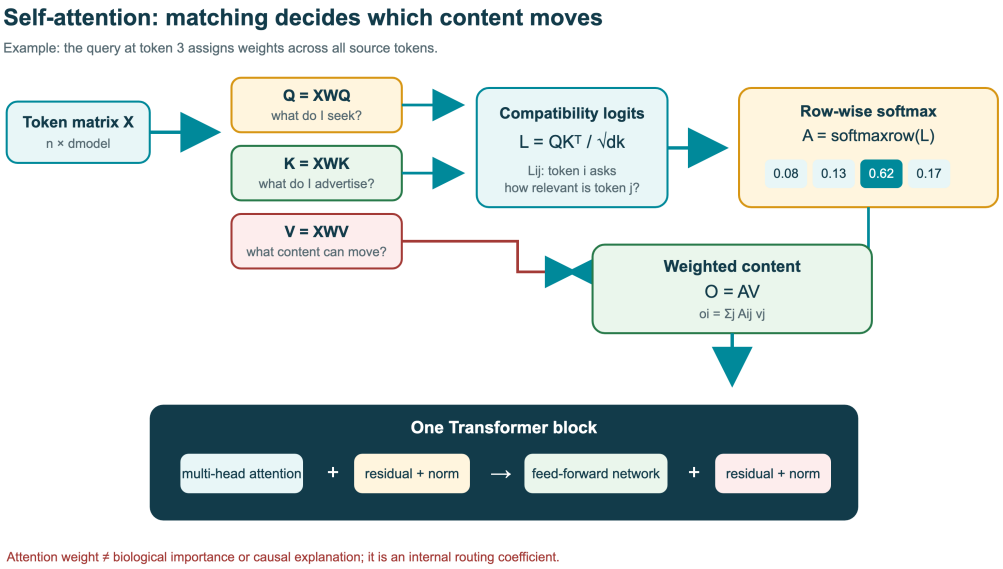


Figure 12.2: Self-attention in a transformer. The token matrix is projected to queries, keys and values; the scaled products of queries and keys, normalised row by row with a softmax, give the weights with which each token mixes the values of all tokens. A transformer block adds residual connections, normalisation and a feed-forward network. The attention weights are internal coefficients of the computation, not by themselves explanations. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

12.4.3 Language models for proteins and DNA

Language models are trained, self-supervised, to predict missing or next tokens in sequences. For proteins, the tokens are amino acids. Two training objectives dominate:

- **masked-token prediction** (as in BERT): a fraction of the tokens, about 15%, is hidden and predicted from the context on both sides. ESM-2, ProtBert, DNABERT and the Nucleotide Transformer are trained this way (Lin et al. 2023; Elnaggar et al. 2022; Ji et al. 2021; Dalla-Torre et al. 2024). Such models are mainly used to produce embeddings and to score mutations;
- **next-token prediction** (autoregressive, as in GPT): each token is predicted from the tokens before it. ProGen, ProtGPT2 and Evo are trained this way (Madani et al. 2023; Ferruz et al. 2022; Nguyen et al. 2024), and can generate new sequences one token at a time (Section 12.5).

In both cases, the labels come from the sequences themselves, which is what makes it possible to train on all known sequences. Models such as ESM (Rives et al. 2021; Lin et al. 2023) and ProtTrans (Elnaggar et al. 2022), trained on hundreds of millions of sequences from UniProt and metagenomes, learn **embeddings**, vector representations of each residue and each protein, that capture structure, function and evolutionary constraints without ever being shown them. Their history and architecture are described in Section 1.10, and their use for structure prediction in Section 9.9.

The embeddings enable **transfer learning**: a large model is pre-trained once on unlabelled sequences, and its representations are then used as input for many downstream tasks with little labelled data, such as localisation (DeepLoc 2, SignalP 6; 3), variant effects (Section 10.6) or function (Section 10.4). A small

model on top of the embeddings, or *fine-tuning* of the whole model, often matches or beats specialised methods.

For **DNA**, language models face longer sequences and a sparser signal, since most of the genome is not under strong constraint (Section 6.9). DNABERT (Ji et al. 2021) and the Nucleotide Transformer (Dalla-Torre et al. 2024) adapted the transformer to DNA, and Evo and Evo 2, trained on genomes from all domains of life, model sequences of up to a million bases and can generate new genomic sequences (Nguyen et al. 2024; Brixi et al. 2025). A complementary approach trains supervised models to predict experimental measurements, such as gene expression, chromatin accessibility and splicing, directly from long DNA sequences: Enformer (Avsec et al. 2021) and AlphaGenome (Avsec et al. 2026) use this to predict the effects of non-coding variants.

12.4.4 Interpretability and reporting standards

Deep networks are often described as *black boxes*. Several methods try to explain what they have learned (Novakovsky et al. 2022): **attribution** methods, such as saliency maps and *in silico* mutagenesis, ask which input positions influence a prediction most, and can reveal the motifs a genomics model has learned; attention weights can be inspected, for example showing that protein language models attend to residues in contact (Rao et al. 2020); and models can be probed with designed inputs. Attention weights are not by themselves explanations: a position that receives high attention need not be the one that changes the prediction, and showing that an input causes an output requires further analysis, such as perturbing the input and measuring the effect. Explanations are useful for finding errors and generating hypotheses, but they are themselves approximations and should not be over-interpreted.

As ML spread through biology, the quality of studies varied widely. Common problems were inadequate test sets, data leakage, missing baselines and code or data that were not available. Recommendations have been proposed to improve this (Jones 2019). The **DOME** recommendations, from the ELIXIR machine learning focus group, ask authors to report four aspects of a supervised ML study: **D**ata (sources, size, splits, redundancy between them), **O**ptimisation (algorithm, hyperparameters, how overfitting was prevented), **M**odel (availability, interpretability) and **E**valuation (metrics, comparison with baselines and other methods, confidence) (Walsh et al. 2021). They are a useful checklist both for designing studies and for reading them critically, which is what the exam expects.

12.4.5 Feature importance and attribution

The same caution applies to the importance measures used for any model, including linear models and random forests.

- **Coefficients** of a linear model depend on the scaling of the features and on which other features are included (4).
- **Impurity-based importance** in random forests favours features with many possible split points, such as continuous features.
- **Permutation importance** measures how much the performance drops when one feature is shuffled. When two features are correlated, the model can use either, and shuffling one of them may change little, so both can appear unimportant.
- **SHAP values** distribute a single prediction among the features, using Shapley values from game theory, relative to a chosen *background* set of examples. **Integrated gradients** sum the gradients of the output along a path from a reference input (for example an all-zero or shuffled sequence) to the actual input.

All of these describe how a particular model makes its predictions. The result depends on the reference or background, on how the features are represented and on how they are correlated, and should be checked for stability across resamples. None of them estimates the effect of changing the feature in a real cell or patient. If a model of drug response gives the largest SHAP value to an immune gene, this is a hypothesis about the gene, to be tested in independent data and by perturbation, not evidence that the gene causes the response. An explanation that agrees with known biology may still reflect a confounder, such as the hospital or batch.

See also: further explanation in Section D.12.3; related Wikipedia articles in Section E.12.3.

i Sources

Parts of this section are adapted from the Wikipedia articles Deep learning, Convolutional neural network, Long short-term memory, Transformer (deep learning architecture), Graph neural network and Explainable artificial intelligence (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Statistics and Machine Learning for Biological Data" (CC BY 4.0), used with permission.

12.5 Generative models

Most of the models above are **discriminative**: they predict a label y for an input $\mathbf{x}$, that is, they model $P(y \mid \mathbf{x})$. **Generative models** instead learn an approximation of the distribution of the data themselves, $P(\mathbf{x})$, so that they can generate new examples, such as new protein sequences, structures or cells. Profile HMMs are an old example: they are generative models of protein families (Section 12.3). Modern generative models use deep networks and come in several families.

- **Autoregressive models** generate a sequence one token at a time, each conditioned on the previous ones, $P(\mathbf{x}) = \prod_i P(x_i \mid x_1, \dots, x_{i-1})$, as in GPT-style language models. Protein language models of this kind, such as ProGen and ProtGPT2, generate new sequences that resemble natural families, and some generated enzymes were active in the laboratory (Madani et al. 2023; Ferruz et al. 2022).
- **Variational autoencoders** (VAEs) compress each example into a low-dimensional *latent* vector with an encoder and reconstruct it with a decoder, trained so that the latent space follows a simple distribution (Kingma and Welling 2014). New examples are generated by sampling the latent space. VAEs of protein families predict the effects of mutations (DeepSequence and EVE, Section 10.6), and scVI models single-cell data (Section 7.6).
- **Generative adversarial networks** (GANs) train a generator to produce examples that a discriminator cannot distinguish from real ones (Goodfellow et al. 2020). They produce sharp images, but are difficult to train and have been less important in biology.
- **Diffusion models** learn to reverse a process that gradually adds noise to the data. Training teaches a network to remove a small amount of noise at each step; generation starts from pure noise and applies the network repeatedly until a clean example emerges (Ho et al. 2020). Diffusion models generate images of very high quality, and in structural biology they generate protein backbones (RFdiffusion, Chroma) and all-atom structures of complexes (AlphaFold3) (Section 9.8, Section 9.12) (Watson et al. 2023; Ingraham et al. 2023; Abramson et al. 2024). *Flow matching* is a closely related, increasingly used approach.

Generative models can be **conditioned** on desired properties: a structure model can be asked for a binder to a given target, a sequence model for a protein with a given function, and multimodal models

such as ESM3 can be prompted with any combination of sequence, structure and function (Hayes et al. 2025). This turns them into design tools.

Evaluating generative models is harder than evaluating predictors, because there is no single correct answer. Generated proteins are assessed for *plausibility* (do they look like natural proteins?), *novelty* (are they different from the training set?), *diversity*, and *designability* (does a structure predictor fold the sequence into the intended structure?). These are all computational proxies. A model that assigns high likelihood to natural sequences, or generates samples that look realistic, has not necessarily learned the true biological distribution, and a plausible sequence is not necessarily a folded, stable or functional protein. In the end, only laboratory experiments show whether a generated protein works, and experimental success rates are the most meaningful measure. Generative models also raise biosecurity questions, since the same tools that design useful proteins could in principle be misused, which has led to discussions of responsible development and screening of DNA synthesis orders.

See also: further explanation in Section D.12.4; related Wikipedia articles in Section E.12.4.

12.6 Foundation models

A **foundation model** is a large model trained on broad data, usually with self-supervision, that can be adapted to many downstream tasks (Bommasani et al. 2021). The term was coined for large language models such as GPT, but the idea, *pre-train once, reuse everywhere*, is now applied to most kinds of biological data. Size and breadth of pre-training alone do not make a model a foundation model. The claim is an operational one, and it should be tested by showing that the model is useful for many tasks it was not trained for, through one of three forms of adaptation:

- **probing**: a simple model, often linear, is trained on frozen embeddings to see what information they contain;
- **fine-tuning**: some or all of the weights are trained further on labelled data for the new task;
- **zero-shot use**: the pre-trained model is used directly, for example by comparing the likelihoods of a wild-type and a mutant sequence to predict the effect of a mutation (Meier et al. 2021).

The self-supervised objectives of Section 12.4, masked-token and next-token prediction, are the usual pre-training tasks. Examples include:

- **proteins**: ESM-2, ESM C and ESM3, ProtTrans and similar language models (Section 9.9);
- **DNA**: the Nucleotide Transformer, Evo and Evo 2 (Dalla-Torre et al. 2024; Brixi et al. 2025);
- **single cells**: Geneformer and scGPT, trained on tens of millions of transcriptomes (Section 7.6) (Theodoris et al. 2023; Cui et al. 2024);
- **structures and complexes**: AlphaFold3 and the co-folding models are sometimes described as foundation models for biomolecular interactions (Section 9.8).

In language, performance improves predictably with model size, data and compute, following **scaling laws** (Kaplan et al. 2020), which motivated ever larger models. In biology the picture is more mixed. Protein language models improve with size up to a point, but gains for many downstream tasks level off, and the amount of truly independent data is limited by evolution: hundreds of millions of sequences represent far fewer independent families.

Foundation models must be evaluated with particular care:

- **Baselines.** Several systematic evaluations have found that foundation models do not beat simple baselines. Single-cell foundation models without fine-tuning were often no better than highly variable genes and PCA (Kedzierska et al. 2025), and deep models did not outperform simple linear models at predicting the effects of genetic perturbations (Ahlmann-Eltze et al. 2025). A

re-evaluation of the single-cell model scBERT found that logistic regression performed as well or better at cell-type annotation, and that removing the pre-training made little difference (Boiarsky et al. 2024), and a benchmark of DNA foundation models found that performance varied between tasks and that specialised models were still better for some, such as predicting gene expression (Feng et al. 2025). A fair comparison gives the baselines the same data and the same effort of tuning, and uses *ablations* (the same architecture without pre-training, or at a smaller scale) to show which part of the model gives the benefit.

- **Benchmark contamination.** Pre-training data are so broad, often all of UniProt, the PDB or large genome collections, that the test sequences or structures of downstream benchmarks, or close homologues of them, have usually been seen during pre-training, although without the labels. This is not ordinary supervised leakage, but it can still invalidate claims that a model generalises to new proteins: the model may only recognise sequences it has seen. Claims of generalisation therefore require test sets that are filtered against the pre-training data as well as the fine-tuning data, or that were created after it (4).
- **Benchmarks.** Standardised benchmarks, such as ProteinGym for the effects of mutations on protein fitness, built from many deep mutational scanning experiments (Notin et al. 2023), make comparisons fairer, but every benchmark can be overfitted by repeated use.
- **Openness and cost.** Training the largest models requires compute available to few groups, and some of the most capable models are closed, so they can be judged only by what their developers report (Section 9.8). Results that cannot be reproduced without such resources are hard to verify. The energy used in training and, for widely used models, in inference has an environmental cost, and the gap in access to compute between large companies, well-funded laboratories and others raises questions of fairness. Smaller models, or distilled versions of large ones, are often sufficient and should be compared.

Learned models usually complement databases and search rather than replace them. AlphaFold2 takes as input a multiple sequence alignment and structural templates found by database searches (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021), and ColabFold made this practical by replacing the slow searches with MMseqs2 (Mirdita et al. 2022). Many predictors combine embeddings with homology transfer by BLAST, and general LLMs are increasingly connected to literature or sequence databases from which they retrieve information before answering (*retrieval-augmented generation*). Such hybrid systems are often more accurate, and easier to update, than a model that must store all knowledge in its weights.

Foundation models are nevertheless changing how bioinformatics is done. Embeddings from a pre-trained model are now a standard starting point for a new predictor, and general-purpose large language models are increasingly used to write code, search the literature and run analyses. They are powerful tools, but, like every method in this book, they must be used with an understanding of what they were trained on, how they were evaluated, and when they fail.

See also: further explanation in Section D.12.5; related Wikipedia articles in Section E.12.5.

Sources

Parts of this section are adapted from the Wikipedia articles Generative model, Variational autoencoder, Generative adversarial network, Diffusion model and Foundation model (CC BY-SA 4.0, retrieved 4 October 2026). Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Statistics and Machine Learning for Biological Data" (CC BY 4.0), used with permission.

12.7 AI in research practice

Artificial intelligence now affects not only the methods of bioinformatics, described in the rest of this chapter, but also the way research is carried out. This section summarises what every researcher is expected to understand, whatever their own topic.

- **Terminology.** *Artificial intelligence* is the broad goal of machines performing tasks that require intelligence. *Machine learning* is the dominant approach: learning from data rather than following explicit rules. *Deep learning* uses neural networks with many layers. *Generative AI* refers to models that produce text, images, sequences or structures, including **large language models** (LLMs) such as those behind ChatGPT, Claude or Gemini. See the terminology box in

    4.

- **How LLMs work, in outline.** They are transformers (Vaswani et al. 2017) trained to predict the next token on very large text collections, and then adapted with instruction tuning and human feedback. You should be able to explain this at the level of Section 1.10, where the same ideas are applied to proteins.
- **Major achievements in biology.** AlphaFold and its successors (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. 2021; Abramson et al. 2024), which were recognised by the 2024 Nobel Prize in Chemistry; protein language models; and generative protein design (Chapter 9). Be ready to discuss what these advances do and do not solve.
- **Using AI tools in research.** LLMs are useful for writing and debugging code, summarising literature and improving language, but they can produce fluent text that is wrong. The main risks are: *invented references* that look real but do not exist or do not support the claim; *fabricated or incorrect facts*, stated with the same confidence as correct ones; *code* that runs but computes the wrong thing, or that is insecure, for example by exposing passwords or installing packages that do not exist or are malicious; *loss of confidentiality* when data are sent to an external service; and *irreproducibility*, since hosted models are updated or withdrawn without notice and the same prompt can give different answers. Check every reference and every claim, test and read generated code, and record the tool, model version, date and prompts used for anything that enters your results. You remain responsible for all output. AI tools cannot be authors of scientific papers (Thorp 2023), and their use should be disclosed according to journal and university policies.
- **Rules and ethics.** Follow Stockholm University's and your department's guidelines on the use of AI tools in courses, examinations and theses. Do not upload unpublished data, confidential information or personal data to external AI services. Be aware of the wider issues: biases inherited from training data, the energy cost of training and using large models, and dual-use concerns, for example in protein design.

See also: further explanation in Section D.12.6; related Wikipedia articles in Section E.12.6.

> **Sources**
>
> Parts of this section are adapted from the Wikipedia articles Artificial intelligence and Large language model (CC BY-SA 4.0, retrieved 4 October 2026).

12.8 Integrating data

No single kind of measurement describes a cell or an organism. DNA sequence constrains what can happen; chromatin, RNA, proteins and metabolites each reflect different processes, with different time scales, noise and missing values. **Multi-omics integration** combines such layers in the hope of answering questions that no single layer can answer (Hasin et al. 2017). Machine learning provides most of the methods, but the main difficulties are about design and interpretation: what is being combined, what question the combination answers, and what it can show. This section draws together ideas from the rest of the chapter and from Section 7.6 and Section 11.3.

12.8.1 What is being integrated?

Before choosing a method, four things must be clear: which **units** were measured (the same cell, the same tissue sample, the same patient, or only comparable cohorts), which **features** the data sets share, what the **goal** is (exploration, finding subtypes, prediction or the effect of an intervention), and to which new data the result should **apply** (another hospital, species or technology). Three regimes are usually distinguished (Figure 12.3):

- **horizontal** integration combines data sets with the same features measured on different units, such as RNA-seq cohorts from several hospitals; the main problem is batch effects;
- **vertical** integration combines different layers measured on the same units, such as RNA and chromatin accessibility in the same cells (Section 7.6), or the genome, transcriptome and proteome of the same tumours (Section 7.7);
- **diagonal** (or *mosaic*) integration combines data sets that share neither all units nor all features, so that the correspondence between them must be inferred, from anchors, prior knowledge or a model.

Paired data, with all layers measured on the same unit, allow the relationship between the layers to be estimated directly. With *unpaired* data the relationship is assumed rather than observed. Mapping identifiers between layers (transcripts to genes, peptides to proteins, metabolite names to database entries; Chapter 2) is a frequent and underestimated source of errors, and collapsing one-to-many mappings can create agreement between layers that is not there.

12.8.2 Early, intermediate and late integration

Methods are often classified by the stage at which the layers are combined.

- **Early integration** concatenates the (normalised) features of all layers into one table and trains a single model. It is simple and works with any method, but a layer with many features or large values can dominate, so scaling and regularisation (4) are essential.
- **Intermediate integration** learns a joint representation, such as shared latent factors, a joint embedding or a fused network, and then uses it for the analysis. It can separate variation shared between layers from variation specific to one, but the latent dimensions are not uniquely defined, and they are harder to interpret than the original features.
- **Late integration** trains one model per layer and combines their predictions, for example by averaging or by a second model (*stacking*). It copes naturally with samples in which a layer is missing, but cannot find interactions between features of different layers. In stacking, the second model must be trained on predictions made for samples that the first-level models were not trained on.

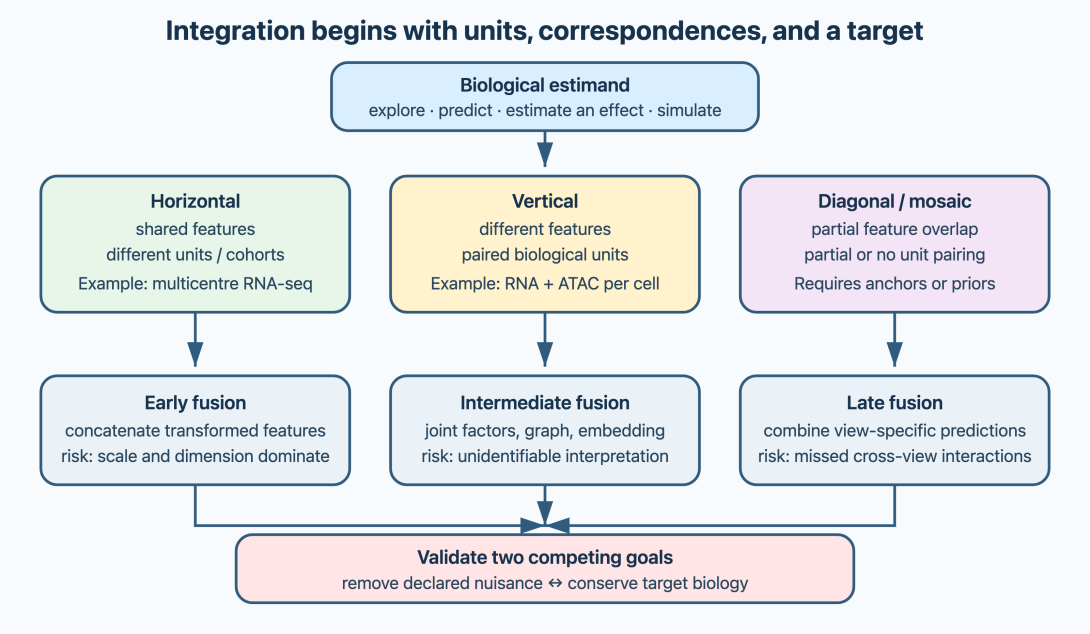


Figure 12.3: Integration starts from the question and from how the data sets correspond. Horizontal integration combines data sets with the same features measured on different units, vertical integration combines different layers measured on the same units, and diagonal integration must infer both correspondences. Early, intermediate and late integration make different assumptions, and every integration must be judged both by how well it removes unwanted variation and by how well it keeps the biology of interest. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

The choice depends on the question and on the data that will be available when the model is used: a predictor that needs proteomics is of little use in a clinic where proteomics is rarely measured. Whatever the method, the gain from integration must be shown by comparison with models of each layer alone and with simple concatenation, evaluated in the same way, with all preprocessing inside the training folds and with splits by patient or site (4). Without these comparisons, an analysis cannot show that integration added anything.

12.8.3 Latent factors and networks

Many integration methods are extensions of the dimensionality reduction used for one data type (Section 7.4).

- **Multi-Omics Factor Analysis** (MOFA, MOFA+) is a multi-layer generalisation of PCA (Argelaguet et al. 2018, 2020). Each layer m is modelled as $X^{(m)} = ZW^{(m)T} + E^{(m)}$: a matrix Z of factor values for each sample, shared by all layers, times a matrix of loadings $W^{(m)}$ specific to each layer, plus noise with a likelihood suited to the layer (Gaussian, count or binary). Priors make the loadings sparse and switch off factors in layers where they are not needed, so that the result shows which sources of variation are shared and which belong to a single layer, and how much variance each factor explains in each layer.
- **Canonical correlation analysis** (CCA) finds combinations of the features of two layers that are maximally correlated with each other; sparse versions are needed when there are more features than samples.
- **Supervised** methods, such as DIABLO, find components that are both correlated between layers and discriminate between known groups of samples (Singh et al. 2019). They are useful for finding

multi-layer biomarker panels, but since the labels shape the components, the result is not an unbiased picture of the data, and the selection must be inside the cross-validation.

- **Similarity network fusion** builds a network of samples for each layer, connecting samples with similar profiles, and fuses them into a single network by repeatedly exchanging neighbourhood information between the layers; clusters of the fused network define, for example, cancer subtypes (B. Wang et al. 2014).
- **Knowledge networks**, such as protein interaction, pathway and functional association networks (Section 11.3), can be combined with measurements, for example by network propagation from the genes that change in an experiment. The meaning of an edge differs between networks (physical binding, regulation, co-expression, shared pathway membership), and well-studied genes have more edges, so results must be compared with randomised networks that keep the degree of each node.

A latent factor is a mathematical summary of covariation, not a biological programme. A factor that is shared by all layers may reflect batch, the hospital, tumour purity or the time a sample spent before freezing as easily as biology, and a factor specific to one layer is not necessarily noise: post-translational regulation, for example, is not expected to be visible in RNA. Factors are interpreted by their loadings, by their association with measured technical and biological variables, by their stability when the data, the number of factors or the preprocessing change, and by whether their association with a phenotype is reproduced in samples that were not used to find them.

Integration across batches and cohorts has the same trade-off as single-cell integration (Section 7.6): removing technical variation can also remove biology, and benchmarks therefore measure both batch removal and the conservation of biological variation (Luecken et al. 2021). If all cases were measured in one batch and all controls in another, disease and batch are completely confounded, and no algorithm can separate them. This can only be prevented by the design of the experiment.

12.8.4 What integration can and cannot show

Integration does not change the basic distinction between **prediction** and **causation** (4). A model that predicts drug response from several omics layers estimates $P(y \mid \mathbf{x})$, the probability of the outcome given the measurements. A causal question asks what would happen to y if a gene, protein or treatment were changed, and correlations in observational data can arise from common causes (*confounders*), from the outcome affecting the measurements, or from how the samples were selected. Agreement between layers strengthens a hypothesis, for example a copy-number gain, higher RNA and protein levels of a receptor, and phosphorylation of its downstream kinase in resistant tumours, but does not show that the receptor causes the resistance: the amplified region may contain other genes, and the phosphorylation may have another cause.

Causal reasoning requires explicit assumptions about how the variables affect each other, often drawn as a *directed acyclic graph*. The graph shows which variables must be adjusted for. Adjusting for a confounder removes bias; adjusting for a *mediator*, a variable on the causal path such as a transcript downstream of a genetic variant, changes the question from the total to the direct effect; and adjusting for a *collider*, a variable affected by both the exposure and the outcome, such as a quality-control measure that depends on both disease and sample quality, creates an association that was not there. Adding every available omics feature as a covariate is therefore not a safe default.

Several designs give stronger causal evidence:

- **Perturbations**, such as CRISPR knockouts or drugs, randomise the change and break confounding. Perturb-seq combines pooled CRISPR screens with single-cell RNA-seq, so that the effect of each perturbation on the whole transcriptome is measured (Dixit et al. 2016) (Section 11.2). Even then,

guide efficiency, off-target effects and the selection of cells matter, and an effect shown in one cell line applies to that context.

- **Mendelian randomisation** uses genetic variants, which are assigned at conception, as natural experiments to estimate the effect of an exposure, such as the level of a protein, on a disease (Davey Smith and Ebrahim 2003). It assumes that the variants affect the disease only through the exposure, an assumption that pleiotropy, linkage disequilibrium and population structure (Section 6.6) can violate.
- **Time courses** and **independent cohorts** help, but temporal order alone does not exclude confounding.

The strength of the evidence must match the strength of the claim (Figure 12.4). That a representation contains information, that a model performs well on a benchmark, and that it works in a new hospital are claims about the model's capability. That a gene mediates a phenotype is a claim about biology, and requires perturbation and independent lines of evidence. A careful conclusion states the narrowest claim that the evidence supports, and the experiment that would test the next step.

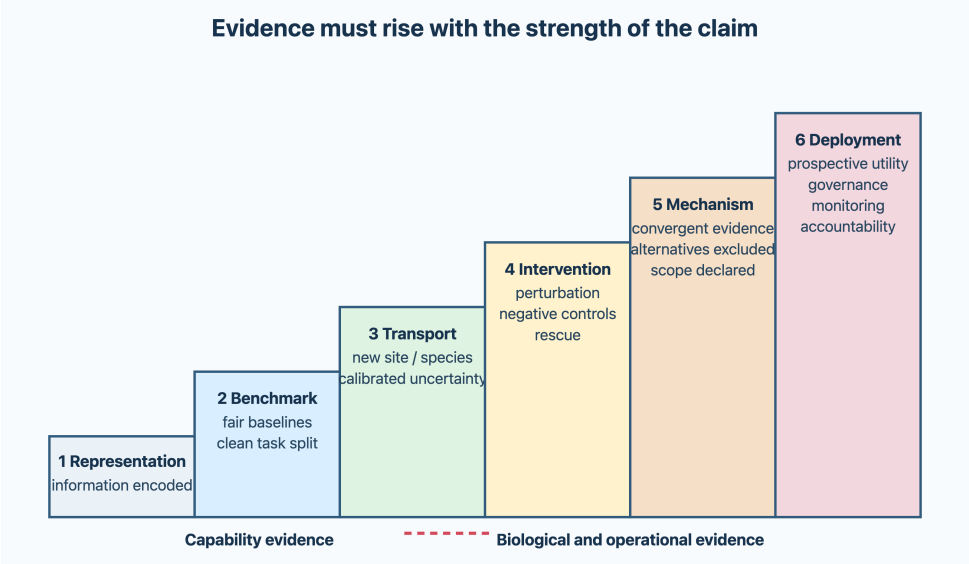


Figure 12.4: The evidence must rise with the strength of the claim. A model's representation, its benchmark performance and its performance on new sites or species are evidence of capability. Claims about interventions and mechanisms require perturbations, negative controls and convergent evidence, and use in practice requires prospective evaluation and monitoring. Figure: Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0).

See also: further explanation in Section D.12.7; related Wikipedia articles in Section E.12.7.

> **Sources**
>
> Parts of this section are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers*, chapter "Integration, Reproducibility, and Emerging Bioinformatics" (CC BY 4.0), used with permission.

12.9 Key concepts

A checklist for revision.

- **Terminology:** AI, ML, deep learning, generative AI and foundation models overlap but are not synonyms
- **Kinds of learning:** supervised, unsupervised and self-supervised; classification vs. regression; loss functions; gradient descent, learning rate, stochastic gradient descent and Adam
- **Linear models:** linear and logistic regression; coefficients as conditional associations; odds ratios; L2 (ridge) shrinks, L1 (lasso) sets weights to zero and selects features, elastic net; feature scaling; unstable selection of correlated features
- **Generalisation:** overfitting and underfitting; bias–variance trade-off; training, validation and test sets (a test set used for decisions becomes a validation set); cross-validation and nested cross-validation for hyperparameter selection; regularisation (L1, L2, early stopping, dropout)
- **Pitfalls of biological data:** non-independent examples and why random splits overestimate performance; homology between training and test sets; redundancy reduction and grouped splits by clusters, folds, patients, donors, chromosomes or dates; data leakage, including preprocessing and feature selection outside the training folds; homology-aware partitioning (GraphPart); class imbalance; biased labels and missing negatives; distribution shift; prediction vs. causation
- **Evaluation:** confusion matrix; accuracy (and why it misleads for imbalanced data), sensitivity, specificity, precision, F1; false-positive rate vs. false discovery rate; MCC; ROC and AUC; precision–recall curves and their baseline; operating points; ranking vs. calibration, reliability diagrams; Brier score and log loss; choosing a threshold from the costs of errors; aleatoric vs. epistemic uncertainty; correlation for regression; benchmark design (retrospective, cluster and temporal splits, external validation, prospective/blind); blind benchmarks (CASP, CAFA, CAGI, CAMI, DREAM); simple baselines
- **History:** the perceptron and its limits (XOR); backpropagation; neural networks for secondary structure and the importance of representation (PHD); HMMs (generative models, and why they belong in this history) and the Viterbi, forward–backward and Baum–Welch algorithms; applications of HMMs; SVMs and kernels; decision trees, random forests and gradient boosting
- **Deep learning:** why depth became possible (data, GPUs, ReLU, residual connections); CNNs and their use for DNA motifs and contact maps; RNNs and LSTMs; attention and transformers (and the limits of long-range modelling); graph neural networks; vanishing gradients; queries, keys and values in self-attention, positional information and the quadratic cost
- **Language models:** self-supervised pre-training; masked-token vs. next-token prediction; embeddings; transfer learning and fine-tuning; protein language models; DNA language models (DNABERT, Nucleotide Transformer, Evo); sequence-to-function models (Enformer, AlphaGenome)
- **Interpretability and reporting:** attribution, in silico mutagenesis, attention maps (and why attention is not explanation); coefficients, permutation importance, SHAP and integrated gradients describe the model, not causal effects; the DOME recommendations (data, optimisation, model, evaluation)
- **Generative models:** discriminative vs. generative; autoregressive models (ProGen, ProtGPT2); VAEs; GANs; diffusion models; conditioning; evaluating generated proteins (plausibility is not function); biosecurity
- **AI in research practice:** AI, ML, deep learning and generative AI; how LLMs are trained; AlphaFold and the 2024 Nobel Prize; using AI tools responsibly (invented references, fabricated facts, code security, data confidentiality, model versions and reproducibility, authorship, disclosure); bias, energy use and dual use
- **Foundation models:** operational definition; probing, fine-tuning and zero-shot use; examples for proteins, DNA, single cells and structures; scaling laws and their limits in biology; evaluation against baselines with equal tuning and ablations, benchmark contamination by pre-training data, benchmarks (ProteinGym); models combined with MSAs, templates and database search; open vs. closed models, compute, energy and access
- **Integrating data:** units, features, goal and transport; horizontal, vertical and diagonal integration;

paired vs. unpaired data; identifier mapping; early, intermediate and late integration and the need for single-layer baselines; MOFA, CCA, DIABLO, similarity network fusion, network propagation; latent factors are not programmes; batch vs. biology and complete confounding; prediction vs. causation; confounders, mediators and colliders; perturbations, Perturb-seq and Mendelian randomisation; matching the evidence to the claim

12.10 Recommended reading

These papers are recommended for reading in full:

- Sapoval et al. (2022) — current progress and open challenges for deep learning across the biosciences (the basis of earlier exams).
- Greener et al. (2021) — a guide to machine learning for biologists; how to choose and evaluate methods.
- LeCun et al. (2015) — deep learning, by three of its founders.
- Krogh (2008) — what artificial neural networks are, in three pages.
- Walsh et al. (2021) — DOME; what a supervised machine learning study in biology should report.
- Chicco and Jurman (2020) — why the Matthews correlation coefficient is better than F1 and accuracy.

Titles, abstracts, links and the full texts are in Section A.12; a longer list of further papers is in Section B.12.

12.11 Background reading (open access)

- **Machine learning (Wikipedia)** — Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without being explicitly programmed.
  https://en.wikipedia.org/wiki/Machine_learning
- **Deep learning (Wikipedia)** — In machine learning, deep learning (DL) focuses on utilizing multilayered neural networks to perform tasks such as classification, regression, and representation learning.
  https://en.wikipedia.org/wiki/Deep_learning
- **Hidden Markov model (Wikipedia)** — In probability theory, a hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process. An HMM requires that there be an observable process whose outcomes depend on the outcomes of in a known way.
  https://en.wikipedia.org/wiki/Hidden_Markov_model
- **Artificial neural network (Wikipedia)** — In machine learning, a neural network (NN) or artificial neural network (ANN) is a computational model inspired by the structure and functions of biological neural networks.
  https://en.wikipedia.org/wiki/Neural_network_(machine_learning)
- **Backpropagation (Wikipedia)** — In machine learning, backpropagation is a gradient computation method commonly used for training a neural network in computing parameter updates.
  https://en.wikipedia.org/wiki/Backpropagation
- **Support vector machine (Wikipedia)** — In machine learning, a support vector machine (SVM) or support vector network is a supervised max-margin model with associated learning algorithms that analyze data for classification and regression analysis.
  https://en.wikipedia.org/wiki/Support_vector_machine

- **Random forest (Wikipedia)** — Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training. For classification tasks, the output of the random forest is the class selected by most trees.
  https://en.wikipedia.org/wiki/Random_forest
- **Overfitting (Wikipedia)** — In mathematical modeling, overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit to additional data or predict future observations reliably.
  https://en.wikipedia.org/wiki/Overfitting
- **Regularization (Wikipedia)** — In mathematics, statistics, finance, and computer science, particularly in machine learning and inverse problems, regularization is a process that converts the answer to a problem to a simpler one. It is often used in solving ill-posed problems or to prevent overfitting.
  https://en.wikipedia.org/wiki/Regularization_(mathematics)
- **Cross-validation (Wikipedia)** — Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
  https://en.wikipedia.org/wiki/Cross-validation_(statistics)
- **Transformer (deep learning architecture) (Wikipedia)** — In deep learning, the transformer is a family of artificial neural network architectures based on the multi-head attention mechanism, in which input data such as text, images, or audio is converted to a sequence of numerical representations called tokens, and each token is converted into a vector through an embedding layer.
  https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)
- **Diffusion model (Wikipedia)** — In machine learning, diffusion models, also known as diffusion-based generative models or score-based generative models, are a class of latent variable generative models. A diffusion model consists of two major components: the forward diffusion process, and the reverse sampling process.
  https://en.wikipedia.org/wiki/Diffusion_model
- **Foundation model (Wikipedia)** — In artificial intelligence, a foundation model (FM), also known as large x model, is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases. Generative AI applications like large language models (LLMs) are common examples of foundation models.
  https://en.wikipedia.org/wiki/Foundation_model

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

12.12 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

12.13 Review problems

These problems are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission. For a longer treatment of this topic, see the chapters "Statistics and Machine Learning for Biological Data" and "Integration, Reproducibility, and Emerging Bioinformatics"

of that book. In the HTML version, click a box to show the extensions, solution outline and common errors.

12.13.1 L1 and L2 regularisation

Compare L1, L2 and elastic-net regularisation for a classifier of patients trained on the expression levels of 20,000 genes, many of which are correlated.

Extensions, solution outline and common errors

Extensions:

- Why do L1 penalties give exactly zero weights, and L2 penalties not?
- Why must the features be scaled before fitting?
- How should the penalty strength λ be chosen?
- How stable is the set of selected genes, and what does it mean biologically?

Solution outline:

1. With a loss $L(\beta)$, L1 (lasso) minimises $L + \lambda \sum_j |\beta_j|$. The penalty has corners on the coordinate axes, so many weights become exactly zero and the model selects features.
2. L2 (ridge) minimises $L + \lambda \sum_j \beta_j^2$. It shrinks all weights smoothly towards zero but rarely sets any to zero, and stabilises the fit when features are correlated.
3. The elastic net combines both, and tends to keep or drop correlated genes together.
4. The penalties act on the size of the weights, so features must be on comparable scales.
5. λ is a hyperparameter, chosen by cross-validation inside the training data (nested cross-validation if performance is also estimated).
6. Selection should be checked by resampling. Selected genes are predictive within this model, not causal: a correlated gene could often replace a selected one with the same performance.

Common errors:

- Reversing L1 and L2.
- Choosing λ on the final test set.
- Treating a sparse signature as unique when correlated genes could replace it.
- Fitting without scaling the features.

Connections: 4 (linear models, regularisation, nested cross-validation); Section 7.4 (correlated expression and differential expression).

12.13.2 Leakage-free evaluation for protein sequences

Design a benchmark for a method that predicts enzyme function from 100,000 protein sequences with family labels.

Extensions, solution outline and common errors

Extensions:

- Compare random, homology-aware, family-held-out and clade-held-out splits.
- Where do normalisation, feature selection, hyperparameter tuning, calibration and the

choice of threshold belong in a nested design?
- What if the method uses embeddings from a protein language model pre-trained on UniProt?

Solution outline:

1. Define the intended novelty: a new member of a known family, a new family, or a new clade. The split follows from this.
2. Cluster sequences by similarity (for example with MMseqs2, or with a partitioning method such as GraphPart) and keep each cluster, and where relevant each family or clade, in one subset.
3. Fix the outer test set first. Fit every learned step and choose the model, features, hyperparameters, calibration and threshold using only the inner training and validation data.
4. Record the pre-trained model and the version of its training data, estimate how many test sequences or close homologues were seen in pre-training, and repeat with stricter filtering.
5. Compare with nearest-neighbour transfer by BLAST, profile methods and a linear model on the embeddings, with the same tuning effort.
6. Report performance both for close homologues and for new families if both claims are made.

Common errors:

- Treating different accessions as independent examples.
- Fitting normalisation or feature selection before splitting.
- Choosing the architecture by test-set performance.
- Ignoring overlap with pre-training data.
- Comparing a heavily tuned model with an untuned baseline.

Connections: 4 (pitfalls and benchmark design); Section 12.6 (benchmark contamination); Section 1.11 (benchmarks of homology search); Section 10.7 (function prediction benchmarks).

12.13.3 Evaluation under class imbalance and distribution shift

A cancer-screening model reports 98% accuracy and an AUC of 0.91 in a cohort with 1% disease prevalence. Does the evidence support clinical use?

Extensions, solution outline and common errors

Extensions:

- Compare the ROC curve with the precision–recall curve.
- Define calibration and the Brier score.
- How should the threshold be chosen?
- What do case–control sampling, external validation and subgroups change?

Solution outline:

1. At 1% prevalence, always predicting “healthy” gives 99% accuracy, so 98% accuracy says nothing on its own.
2. The AUC measures ranking of positives above negatives. Precision–recall curves focus on the positive predictions; their random baseline equals the prevalence, so they show whether the top predictions are correct.
3. Whether the scores are probabilities requires a reliability diagram and a proper score such as the Brier score or log loss.

4. The threshold is chosen on validation data from the costs of false negatives and false positives.
5. Performance needs confidence intervals and evaluation in external sites, later time periods and relevant subgroups.
6. If the prevalence in a case–control sample differs from the screening population, the probabilities must be recalibrated.

Common errors:

- Treating the AUC as sufficient because it does not depend on prevalence.
- Choosing the threshold on the final test cohort.
- Reporting subgroup results without uncertainty.
- Assuming that retrospective discrimination shows prospective usefulness.

Connections: 4 (evaluation, calibration, distribution shift); Section 10.6 (variant effect prediction).

12.13.4 Diagnosing a violated assumption

A classifier that performs very well in random cross-validation fails on samples from a new laboratory. Give at least four explanations and design diagnostics to distinguish them.

Extensions, solution outline and common errors

Extensions:

- How can a negative control feature reveal leakage?
- Which kinds of distribution shift can recalibration fix, and which need retraining or a new design?
- How can you tell whether the model used a signal correlated with the laboratory?

Solution outline:

1. Possible explanations: batch artefacts specific to the laboratory; leakage through preprocessing or identifiers; a shift in the distribution of the features; a different prevalence; a different relation between features and labels; inconsistent labelling; and biology not covered by the training data.
2. Plot features and predictions by laboratory and by biological covariates, and check missing values, controls and instrument metrics.
3. Repeat the evaluation with splits by laboratory, and with all preprocessing fitted inside the training folds.
4. Test whether the laboratory can be predicted from the features, and whether predictions remain associated with laboratory after accounting for the true label.
5. Recalibration can correct a change in prevalence or in the scale of the scores, but not missing biology or a confounded training design.

Common errors:

- Batch-correcting the whole data set before splitting.
- Treating distribution shift as a single phenomenon.
- Reporting only overall accuracy.
- Interpreting laboratory-specific features as disease biology.

Connections: 4 (data leakage, distribution shift); Section 13.6 (site signatures in pathology images); Section 7.4 (batch effects).

12.13.5 Transformers and biological foundation models

Explain how a transformer represents a biological sequence, and what evidence is needed to conclude that large-scale pre-training improves a downstream biological task.

Extensions, solution outline and common errors

Extensions:

- Write down scaled dot-product attention.
- Discuss tokens, positional information, self-supervised objectives and the quadratic cost of attention.
- Which comparisons separate the effect of pre-training from that of the architecture and the model size?

Solution outline:

1. Tokens (residues, bases or k-mers, genes) are embedded and projected to queries, keys and values; attention is $\text{softmax}(QK^T/\sqrt{d_k})V$, and several heads use different projections.
2. Attention ignores order, so positional information is added. The cost grows with the square of the sequence length.
3. Pre-training uses masked-token or next-token prediction; the embeddings reflect this objective and the training data.
4. A benefit of pre-training is shown only with leakage-aware splits, baselines with the same data and tuning, ablations (the same model without pre-training or at smaller size), external data sets, calibration and, for claims about function, experiments.
5. An attention weight is not in itself an interaction or an explanation.

Common errors:

- Treating attention as a mechanistic explanation.
- Calling any large network a foundation model.
- Assuming that a high-likelihood generated sequence folds, functions or is safe.
- Comparing a heavily optimised model with an untuned baseline.

Connections: Section 12.4 (transformers and language models); Section 12.6 (evaluating foundation models); Section 1.10 and Section 9.9 (protein language models).

12.13.6 Interpreting feature attributions

A model that predicts response to therapy from tumour expression data gives the largest SHAP value to an immune gene. What can be concluded, and what would be needed before the model is used in the clinic?

Extensions, solution outline and common errors

Extensions:

- How do correlated features and the choice of background affect SHAP values?
- How would you check that the attribution is stable?
- What evidence would make the gene a candidate therapeutic target?

Solution outline:

1. SHAP distributes this model's prediction among the features relative to a background set; it says how the model predicts, not what the gene does.
2. Correlated genes can share or exchange attribution, so stability must be checked across folds, sites, backgrounds and preprocessing choices.
3. The association might reflect a confounder, such as hospital, batch or tumour purity.
4. A causal role requires independent cohorts and perturbation experiments.
5. Clinical use requires a defined population, external or prospective validation, calibrated probabilities, analysis by site and subgroup, fixed versions of data and model, and monitoring after deployment.

Common errors:

- Turning a feature attribution directly into a drug target.
- Assuming interpretability repairs leakage or confounding.
- Treating predictive performance as causal understanding.

Connections: Section 12.4 (interpretability and feature attribution); Section 12.8 (prediction and causation); Section 7.7 (cancer genomics).

12.13.7 Designing a trustworthy method benchmark

A new method reports the highest average score across ten public data sets. How strong is this evidence, and how would you design a more informative benchmark?

Extensions, solution outline and common errors

Extensions:

- Which kinds of leakage are specific to biology?
- Why do public benchmarks lose their value over time?
- Which baselines and measures are required, and what would count as a meaningful advance?

Solution outline:

1. Restate the task, the population the method will be used on, the independent unit and what counts as new.
2. Check for overlap of patients, donors, families, homologues, sites, time periods and near-duplicates across splits.
3. Keep normalisation, imputation, feature selection and tuning inside the training data, and check pre-training data for overlap with the test sets.
4. Compare with simple, specialised and earlier methods, with ablations, using the same data and tuning effort.

5. Report results per data set with uncertainty, not only the mean rank, together with calibration, failures and compute.
6. Prefer blind or prospective tests, and an improvement larger than the variation caused by analysis choices.

Common errors:

- Selecting the best random seed or the most favourable measure.
- Calling a small, statistically detectable gain a breakthrough.
- Assuming a public test set stays uncontaminated.

Connections: 4 (benchmark design and blind benchmarks); Section 9.2 (CASP); Section 12.6 (contamination).

12.13.8 Defining multi-omics integration

Genome, RNA, protein and metabolite measurements are available for a cohort of patients. Propose an integration strategy, and state when integration would not be justified.

Extensions, solution outline and common errors

Extensions:

- Are the layers measured on the same units?
- Compare horizontal, vertical and diagonal integration.
- Compare early, intermediate and late integration.

Solution outline:

1. Define the goal: harmonisation, exploration, subtypes, prediction or mechanism.
2. Map each layer to patient, tissue sample, cell and time point; distinguish paired from unpaired data.
3. Do quality control for each layer before combining them, and keep one-to-many identifier mappings.
4. Explain early concatenation, joint latent representations and late combination of predictions, and how each handles missing layers.
5. Fit single-layer and simple concatenation baselines with the same validation design.
6. Require independent evidence before interpreting a shared factor biologically.
7. Do not integrate when a single layer already answers the question, or when complete confounding makes the question impossible to answer.

Common errors:

- Assuming more layers automatically means more information.
- Calling the largest latent factor a biological programme without checking batch.
- Analysing only complete cases and hiding why layers are missing.

Connections: Section 12.8 (integrating data); Section 7.6 (multimodal single-cell data); Chapter 2 (identifiers).

12.13.9 Interpreting a latent factor

MOFA finds a latent factor that is associated with disease and has large loadings in RNA, methylation and protein data. What can be concluded, and how would you validate it?

Extensions, solution outline and common errors

Extensions:

- Write down the factor model.
- How do shared and layer-specific factors differ?
- Is the factor uniquely defined?

Solution outline:

1. The model is $X^{(m)} = ZW^{(m)T} + E^{(m)}$, with factor values Z shared by all layers, loadings $W^{(m)}$ for each layer, and a likelihood suited to each layer.
2. The factor summarises covariation; it is not an observed mechanism.
3. Check the variance explained per layer, the stability of the loadings, association with technical variables, missing data, the number of factors, preprocessing and random seed.
4. The sign and order of factors are arbitrary; match factors between runs by their loadings.
5. Test the association with disease in held-out patients.
6. Use independent assays or perturbations to test a proposed pathway.

Common errors:

- Interpreting a two-dimensional plot without looking at the original measurements.
- Treating a layer-specific factor as noise.
- Using the same outcome to find and to validate the factor.
- Inferring causal direction from correlated loadings.

Connections: Section 12.8 (latent factors); Section 7.4 (PCA and dimensionality reduction).

12.13.10 Distinguishing prediction from causal mechanism

A multi-omics classifier predicts drug response accurately in held-out samples. Does it identify a mechanism of resistance? What design and evidence would a causal claim need?

Extensions, solution outline and common errors

Extensions:

- What is the difference between $P(y \mid \mathbf{x})$ and the effect of an intervention?
- How should a causal graph be used?
- Why might adjusting for expression introduce bias?
- What would Perturb-seq add?

Solution outline:

1. Prediction estimates performance on new cases from a population; a mechanism concerns what changes under intervention.
2. Define treatment, outcome, population, time and the effect of interest (total or direct).

3. Draw plausible confounders, mediators, colliders and selection processes before choosing covariates.
4. Adjusting for a mediator changes the question; adjusting for a collider can create bias.
5. Use genetic or pharmacological perturbations with negative controls, rescue, time courses and several genetic backgrounds.
6. Perturb-seq measures the transcriptome-wide effect of each perturbation, but guide efficiency, off-target effects and cell selection must be checked.

Common errors:

- Treating feature importance as a causal effect.
- Assuming that temporal order removes confounding.
- Adjusting for every available omics feature.
- Calling an in silico knockout an experiment.

Connections: Section 12.8 (what integration can and cannot show); Section 11.2 (Perturb-seq and perturbation screens); Section 12.4 (feature attribution).

13 Image analysis and models of cells

> **Not in the exam description**
>
> This topic is not part of the 2025 exam description. It was added in version 0.4 of this book because image-based methods, and the large projects that aim to classify cell types and build models of whole cells, have become a central part of bioinformatics. Ask your examiner whether it is included in your exam.

13.1 Overview

Images are now one of the largest sources of biological data. Automated microscopes image millions of cells in a single screen, histology slides are scanned at high resolution in hospitals, and electron tomography shows the molecules inside cells. Deep learning has made it possible to analyse such images at scale, and the results are increasingly combined with sequence, expression and structure data. This chapter builds on the machine-learning methods of Chapter 12 and covers:

- images as data: kinds of images, classical analysis, software, file formats and repositories (Section 13.2);
- deep learning for images: classification and segmentation, U-Net, Cellpose and StarDist, vision transformers, self-supervised learning and foundation models, and the pitfalls (Section 13.3);
- image-based profiling: Cell Painting, morphological profiles and the JUMP and RxRx data sets (Section 13.4);
- protein localisation from images: the subcellular atlas of the Human Protein Atlas, image-classification challenges, OpenCell and self-supervised models (Section 13.5);
- tissues: multiplexed imaging, spatial omics and computational pathology (Section 13.6);
- cell atlases and AI virtual cells: the Human Cell Atlas, HuBMAP, the virtual-cell efforts of CZI, the Arc Institute and SciLifeLab's Alpha Cell, and how well current models work (Section 13.7);
- mechanistic and structural models of whole cells, and cryo-electron tomography (Section 13.8).

The chapter connects to many others: spatial proteomics and the Human Protein Atlas in Chapter 8, sequence-based localisation prediction in Chapter 10, single-cell and spatial transcriptomics in Chapter 7, and cryo-EM and molecular dynamics in Chapter 9.

13.2 Images as data

Microscopy is the oldest way of studying cells, and for most of its history the results were judged by eye. Modern microscopes produce far more images than anyone can inspect: an automated screen can image millions of cells, and a single light-sheet experiment can produce terabytes. **Bioimage analysis** turns these images into measurements that can be analysed with the same statistical and machine-learning methods as other omics data.

13.2.1 What an image is

A digital image is an array of numbers. A fluorescence image has one *channel* for each fluorescent label, and may in addition have several focal planes (z) and time points (t), so that a single experiment is a five- dimensional array (x, y, z, channel, time). Each value is an intensity, not a count of molecules: it depends on the label, the microscope, the exposure and the detector, and images are only comparable when they were acquired in the same way. The main kinds of images in biology are:

- **Transmitted-light** (brightfield, phase contrast) images, which need no labels but show little molecular detail;
- **fluorescence** images (widefield, confocal, light-sheet and super-resolution microscopy), where proteins or structures are made visible with antibodies, dyes or fluorescent proteins;
- **electron microscopy**, which reaches molecular resolution but only for fixed or frozen samples (Section 13.8);
- **histology**, thin tissue sections stained, for example with haematoxylin and eosin (H&E), and scanned as *whole-slide images* of up to 100,000 × 100,000 pixels (Section 13.6).

13.2.2 Classical image analysis

The classical workflow, still used in many pipelines, has four steps: *preprocessing* (background and illumination correction, denoising), *segmentation* (finding the objects, usually nuclei and cells, for example by thresholding followed by a watershed transform to split touching objects), *feature extraction* (measuring size, shape, intensity and texture for each object) and *analysis* of the resulting table. Each step involves choices, and the results depend on them; this is why image analysis is often less reproducible than it appears.

13.2.3 Tools, formats and repositories

Most bioimage analysis uses open-source software. **ImageJ** and its distribution **Fiji** are the general-purpose tools of light microscopy, extended by thousands of plugins (Schindelin et al. 2012; Rueden et al. 2017). **CellProfiler** builds analysis pipelines from modules without programming and is the standard tool for high-throughput screens (Carpenter et al. 2006; Stirling et al. 2021). **ilastik** trains pixel and object classifiers interactively from a few user annotations (Berg et al. 2019). **QuPath** is designed for whole-slide pathology images (Bankhead et al. 2017), and **napari** is a multidimensional image viewer for Python, to which many deep-learning tools are added as plugins.

Images need metadata (organism, sample preparation, microscope and acquisition settings) to be reusable. The REMBI recommendations define the minimum metadata for publishing biological images (Sarkans et al. 2021), and the OME-Zarr format (OME-NGFF) stores large multidimensional images in chunks that can be read directly from the cloud (Moore et al. 2021, 2023). Published images are deposited in the **BioImage Archive** at EMBL-EBI (Hartley et al. 2022), and curated reference data sets, linked to genes, compounds and phenotypes, in the **Image Data Resource** (IDR) (Williams et al. 2017). The **Broad Bioimage Benchmark Collection** (BBBC) provides annotated image sets for testing methods (Ljosa et al. 2012), and the **Cell Tracking Challenge** has benchmarked segmentation and tracking methods for more than ten years (Maška et al. 2023). The volume of data is part of the challenge: unlike sequences, images cannot be compressed to a few bytes per measurement, and moving them is often harder than analysing them.

See also: further explanation in Section D.13.1; related Wikipedia articles in Section E.13.1.

Sources

Background from the Wikipedia articles Microscopy, Fluorescence microscope, Image segmentation and Fiji (software) (CC BY-SA 4.0, retrieved 5 October 2026).

13.3 Deep learning for images

Image analysis is where modern deep learning started, and microscopy was one of the first scientific fields to adopt it (Moen et al. 2019). The general methods are described in Section 12.4; this section describes how they are used for biological images.

13.3.1 Tasks

Four tasks cover most applications, and it is worth keeping them apart:

- **classification** assigns a label to a whole image or to a cell (for example a cell type, a cell-cycle phase or the location of a protein);
- **segmentation** assigns each pixel to a class (*semantic segmentation*, e.g. nucleus vs. background) or to an individual object (*instance segmentation*, which separates touching cells);
- **detection and tracking** find objects, such as cells or particles, and follow them over time;
- **image-to-image** methods restore noisy images or predict one kind of image from another, for example fluorescence from brightfield.

13.3.2 Convolutional networks and U-Net

Convolutional neural networks (CNNs) learn filters that detect local patterns and combine them into increasingly abstract features. The breakthrough came in 2012, when AlexNet won the ImageNet challenge by a large margin (Krizhevsky et al. 2017), and residual connections later made networks with more than a hundred layers trainable (He et al. 2016). For biomedical images, the most influential architecture is **U-Net** (Ronneberger et al. 2015). An encoder reduces the image to coarse features, a decoder expands them back to full resolution, and *skip connections* pass fine detail from the encoder to the decoder, so that the network produces precise pixel-wise output from relatively few training images. A U-Net plugin for ImageJ made the method usable without programming (Falk et al. 2018).

13.3.3 Instance segmentation of cells

Separating touching cells is the central problem of cell segmentation. **StarDist** describes each nucleus as a *star-convex polygon*: for every pixel, a CNN predicts whether it belongs to an object and the distances to the object boundary along a fixed set of rays, and overlapping candidates are removed by non-maximum suppression (Schmidt et al. 2018). **Cellpose** predicts, for every pixel, the direction towards the centre of its cell (a *flow field*), and groups pixels whose flows end at the same point. Trained on a diverse collection of images, it segments many kinds of cells without retraining (Stringer et al. 2020), and Cellpose 2.0 adds human-in-the-loop training, so that a model can be adapted to new data from a few corrected images (Pachitariu and Stringer 2022). For tissue images, Mesmer was trained on TissueNet, more than a million manually annotated cells, and reached the accuracy of human annotators (Greenwald et al. 2021).

13.3.4 Transformers, self-supervision and foundation models

Vision transformers (ViTs) split an image into patches and treat them as a sequence of tokens, as language models do with words (Dosovitskiy et al. 2021). They need more training data than CNNs, which has made **self-supervised learning** important. In *DINO*, a student network learns to match the output of a teacher network on different views of the same image, without labels (Caron et al. 2021), and DINOv2 scaled this to general-purpose visual features (Oquab et al. 2023). *Masked autoencoders* (MAE) hide most of the patches of an image and learn to reconstruct them (He et al. 2022), the image counterpart of masked language modelling (Section 12.4). For microscopy, Cell-DINO showed that self-supervised features of fluorescence images classify protein localisation well with few labels (Moutakanni et al. 2025).

The **Segment Anything** model (SAM), trained on more than a billion masks, segments objects from a prompt such as a point or a box (Kirillov et al. 2023). It was not trained on microscopy, but versions fine-tuned for light and electron microscopy, such as µSAM (Archit et al. 2025) and CellSAM (Marks et al. 2025), are now general segmentation tools. These are examples of foundation models (Section 12.6) for images.

13.3.5 Label-free prediction

Networks can also learn to predict one kind of image from another. *In silico labelling* predicted fluorescent labels of nuclei, cell types and dead cells from transmitted-light images of unlabelled cells (Christiansen et al. 2018), and similar models from the Allen Institute predicted the three-dimensional fluorescence of several cellular structures from brightfield images (Ounkomol et al. 2018). Such predictions are useful, but they can only reproduce what is predictable from the input image, and they should not be mistaken for measurements.

13.3.6 Pitfalls

The general pitfalls of machine learning (4) all apply, and some are specific to images. Images from different plates, microscopes, days or hospitals differ systematically (*batch effects*), and a network may learn these differences rather than biology; training and test sets should therefore be split by experiment, not by image. Ground-truth annotations are produced by people, who disagree with each other. Restored or predicted images can contain plausible structures that are not in the sample. Guidelines for validating and reporting deep-learning image analysis address these problems (Laine et al. 2021), and good introductions for biologists are available (Chamier et al. 2019; Hallou et al. 2021). Notebooks such as ZeroCostDL4Mic let users train common models without installing software (Chamier et al. 2021).

See also: further explanation in Section D.13.2; related Wikipedia articles in Section E.13.2.

Sources

Background from the Wikipedia articles Convolutional neural network, U-Net and Vision transformer (CC BY-SA 4.0, retrieved 5 October 2026).

13.4 Image-based profiling

A cell's appearance reflects its state. **Image-based profiling** (also called morphological profiling or *phenomics*) uses this systematically: cells are treated with chemical or genetic perturbations, imaged,

and described by thousands of measurements, so that perturbations can be compared in the same way as expression profiles (Chapter 7).

13.4.1 Cell Painting

The most widely used assay is **Cell Painting** (Gustafsdottir et al. 2013; M.-A. Bray et al. 2016). Six fluorescent dyes, imaged in five channels, label eight cellular components: the nucleus (DNA), nucleoli and cytoplasmic RNA, the endoplasmic reticulum, actin, the Golgi apparatus, the plasma membrane and mitochondria. The assay is cheap, needs no antibodies, and works in many cell types. Software such as CellProfiler segments each cell and measures its size, shape, intensity and texture in every channel, giving more than a thousand features per cell (Section 13.2). A third version of the protocol was optimised quantitatively by the JUMP consortium (Cimini et al. 2023).

13.4.2 From images to profiles

The analysis follows a standard workflow (Caicedo et al. 2017):

1. **Single-cell features** are extracted, either hand-designed features from CellProfiler or embeddings from deep networks (Section 13.3).
2. Features are **aggregated** per well, usually as the median over cells, **normalised** against negative controls on the same plate, and reduced to a smaller set of informative, uncorrelated features.
3. **Batch effects** between plates and laboratories are corrected, the same problem as in expression data (Section 7.4).
4. The profiles are compared: perturbations are tested for whether they change the cells at all, and similar profiles are matched, for example by cosine similarity, to group compounds by mechanism or to connect a gene to a compound that mimics its loss.

Profiles are used to find the mechanism of action of compounds, to identify toxicity, to group genes into pathways, and to find disease phenotypes that a drug reverses.

13.4.3 Large public data sets

The **JUMP Cell Painting Consortium**, a collaboration of the Broad Institute and pharmaceutical companies, has released Cell Painting images of more than 100,000 chemical and genetic perturbations. A pilot data set, CPJUMP1, contains about three million images of cells treated with matched chemical and genetic perturbations of the same genes, designed as a benchmark for methods (Chandrasekaran et al. 2024). A morphological map of the over- and underexpression of more than 15,000 genes in human cells followed (Chandrasekaran et al. 2025). The company Recursion has released RxRx3, images and embeddings of more than 17,000 gene knockouts and 1,674 compounds, with some identities hidden (Fay et al. 2023). These data sets are large enough to train foundation models of cell morphology, and to test whether such models can predict the effects of perturbations that were not measured (Section 13.7).

See also: further explanation in Section D.13.3; related Wikipedia articles in Section E.13.3.

> **Sources**
>
> Background from the Wikipedia article High-content screening (CC BY-SA 4.0, retrieved 5 October 2026).

13.5 Protein localisation from images

Where a protein is in the cell is part of its function. 3 describes how localisation is predicted from sequence; this section describes how it is measured and classified from images, which also provides the training data for the sequence-based predictors.

13.5.1 The subcellular atlas of the Human Protein Atlas

The subcellular section of the Human Protein Atlas (Section 8.6) is based on immunofluorescence: each protein is stained with antibodies in cell lines, together with markers for the nucleus, microtubules and the endoplasmic reticulum, and imaged by confocal microscopy. The first version mapped 12,003 proteins to 30 subcellular structures and showed that about half of all proteins are found in more than one location (Thul et al. 2017). Many proteins also vary between individual cells of the same cell line; about a fifth of the proteome shows such cell-to-cell variability, and only part of it is explained by the cell cycle (Mahdessian et al. 2021).

13.5.2 Image classification as a community challenge

Annotating localisation by eye does not scale, so the HPA organised two open Kaggle competitions. In the first, more than 2,000 teams classified images into 28 location classes, with up to six labels per image. The best models clearly outperformed the earlier automated method, but did not reach the accuracy of human experts (Ouyang et al. 2019). The second competition was *weakly supervised*: labels were given for whole images, and the models had to predict the localisation of each individual cell (Le et al. 2022). The two competitions illustrate several general issues of image classification: multi-label output, rare classes, the choice of evaluation metric (macro-F1 gives each class equal weight) and labels that are themselves uncertain.

13.5.3 Endogenous tagging and self-supervised models

Antibodies can bind other proteins than their target. An alternative is to tag the endogenous gene with a fluorescent label. **OpenCell**, at the Chan Zuckerberg Biohub, tagged more than a thousand human proteins with CRISPR, imaged them in living cells and identified their interaction partners by mass spectrometry; proteins with similar localisation patterns often interact (Cho et al. 2022). Using these images, **cytoself** learned a representation of localisation without any labels, by self-supervised learning, and grouped proteins down to the level of individual complexes (Kobayashi et al. 2022). **SubCell** applies self-supervised vision transformers to the whole HPA image collection and produces a hierarchical map of the proteome learned from images alone (Gupta et al. 2024). Such image embeddings can be combined with sequence embeddings from protein language models (Section 1.10).

13.5.4 Organisation of the whole cell

The **Allen Institute for Cell Science** has taken a complementary approach: instead of many proteins in one cell line, it tagged 25 key structures in human induced pluripotent stem cells and imaged more than 200,000 living cells in three dimensions. The data show that the organisation of the cell's interior is robust to large variation in cell shape, and provide a quantitative description of what a "normal" cell looks like (Viana et al. 2023). The images are available through the Allen Cell Explorer.

See also: further explanation in Section D.13.4; related Wikipedia articles in Section E.13.4.

Sources

Background from the Wikipedia articles Human Protein Atlas and Immunofluorescence (CC BY-SA 4.0, retrieved 5 October 2026).

13.6 Tissues: multiplexed imaging and computational pathology

Cells in a tissue are defined not only by what they express but also by where they are and which cells they touch. Images of tissues are therefore analysed both to classify cell types and to describe how they are arranged.

13.6.1 Multiplexed imaging and spatial omics

Highly multiplexed imaging measures tens of proteins in the same tissue section (Section 8.6). In CODEX, antibodies carry DNA barcodes that are read in repeated cycles (Goltsev et al. 2018). In imaging mass cytometry (Giesen et al. 2014) and multiplexed ion beam imaging (MIBI) (Angelo et al. 2014), antibodies are labelled with metal isotopes that are read by mass spectrometry. Imaging-based spatial transcriptomics, such as MERFISH (Chen et al. 2015) and Xenium (Janesick et al. 2023), measures hundreds to thousands of RNAs in the same way (Section 7.6); an overview of the technologies is given by Park et al. (2022).

The analysis is the same for all of these methods: *segment* the cells (Section 13.3), *quantify* each marker in each cell, *assign cell types*, and analyse *spatial neighbourhoods*, for example which cell types are found next to each other more often than expected by chance. Cell types are usually assigned by clustering, as for single-cell RNA-seq, but they can also be classified directly from the images; CellSighter, a CNN trained on a few thousand labelled cells, reaches about the agreement between human annotators (Amitay et al. 2023).

13.6.2 Computational pathology

Pathologists diagnose cancer and other diseases from H&E-stained tissue sections, and digital scanners have made these slides available for machine learning. A whole-slide image is too large to process at once, so it is split into thousands of small *tiles*, each tile is embedded with a network, and the embeddings are aggregated, often with attention, to make a prediction for the slide (*multiple-instance learning*). Networks trained in this way grade tumours, predict mutations, and predict gene expression from morphology, which links histology to spatial transcriptomics (He et al. 2020).

Since 2024, several **pathology foundation models** have been trained by self-supervised learning on very large collections of slides: UNI on more than 100,000 slides (Chen et al. 2024), Virchow on about 1.5 million (Vorontsov et al. 2024) and Prov-GigaPath on more than 170,000 slides, with a model of the whole slide rather than single tiles (Xu et al. 2024). CONCH was trained on pairs of images and captions, so that slides can be searched and classified with text (Lu et al. 2024). Their embeddings are used as general features for many clinical tasks.

A warning example shows why careful evaluation is needed. In The Cancer Genome Atlas, the hospital that submitted a slide can be recognised from the image itself, from differences in staining and scanning, even after colour normalisation. Models can use these site signatures to predict outcomes, which inflates accuracy when training and test sets contain slides from the same sites (Howard et al. 2021). This is the image version of the data leakage described in 4.

See also: further explanation in Section D.13.5; related Wikipedia articles in Section E.13.5.

Sources

Background from the Wikipedia articles Digital pathology and Spatial transcriptomics (CC BY-SA 4.0, retrieved 5 October 2026).

13.7 Cell atlases and virtual cells

The methods of this chapter and of Section 7.6 are now being combined in very large projects. Their first aim is to catalogue all human cell types; their more ambitious aim is a model that can predict how a cell behaves.

13.7.1 Cell atlases

The **Human Cell Atlas** (HCA), launched in 2016, aims to describe every cell type in the human body by its molecular profile, location and morphology, and to make all data openly available (Regev et al. 2017). A collection of papers published in 2024 described progress from a census of cells towards three-dimensional maps of organs and foundation models trained on the atlas (Rood et al. 2024). The NIH **Human BioMolecular Atlas Program** (HuBMAP) builds three-dimensional maps of human tissues at single-cell resolution, combining imaging with omics (HuBMAP Consortium 2019; Jain et al. 2023). The Human Protein Atlas (Section 8.6) and the Allen Institute for Cell Science (Section 13.5) contribute the protein and imaging dimensions.

13.7.2 The idea of an AI virtual cell

A widely cited perspective from 2024 proposed the **AI virtual cell**: a model, learned from data, that represents molecules, cells and tissues in different states, and can *simulate* the effect of perturbations that have never been measured (Bunne et al. 2024). Unlike the mechanistic models of Section 13.8, such a model would not be built from known reactions but learned from atlases, perturbation screens and images. Its proposed uses include predicting the effects of drugs and mutations and designing experiments in silico.

Several large initiatives are working towards this goal:

- The **Chan Zuckerberg Initiative** (CZI) and its Biohubs have made virtual cells a central aim. They run an open *Virtual Cells Platform* with models, data sets and benchmarks, and in 2025 started the *Billion Cells Project* to generate single-cell data for training models. Models include TranscriptFormer, trained on single-cell data from twelve species (Pearce et al. 2025), and rBio, a language model trained with virtual cell models as a source of feedback (Istrate et al. 2025).
- The **Arc Institute** developed State, a model of the response of cells to perturbations trained on more than 100 million perturbed cells (Adduri et al. 2026), and runs the annual **Virtual Cell Challenge**, a blind test of perturbation prediction in the spirit of CASP (Roohani et al. 2025) (Section 9.2).
- **Alpha Cell**, a strategic programme of SciLifeLab funded by the Knut and Alice Wallenberg Foundation and led by Jan Ellenberg and Mathias Uhlén, aims to build predictive AI models of human cells. It starts from the data of the Human Protein Atlas, and adds spatial models from light and electron microscopy and temporal models of how cells change. In 2026 it was funded until 2033, with KTH, Stockholm University, Karolinska Institutet and Uppsala University as partners (scilifelab.se/alpha-cell).

- Google DeepMind has said publicly that a virtual cell is a long-term goal after AlphaFold, but has not described a specific project.

13.7.3 How good are the models?

Predicting the effect of a perturbation on gene expression is the most common test of a virtual cell. GEARS used a knowledge graph of gene–gene relations to predict the effect of combinations of perturbations (Roohani et al. 2023), and single-cell foundation models (Section 7.6) have been fine-tuned for the same task. Independent evaluations, however, found that deep-learning models did not outperform deliberately simple linear baselines (Ahlmann-Eltze et al. 2025), and that the embeddings of foundation models used without fine-tuning were often no better than simpler methods (Kedzierska et al. 2025). The organisers of the first Virtual Cell Challenge in 2025 reached a similar conclusion: most models did not consistently beat simple baselines across all metrics, and no single metric captured what makes a model good. These results do not show that the goal is wrong, but they show that the claims should be tested, as for protein structure prediction, by blind benchmarks against simple baselines (4).

See also: further explanation in Section D.13.6; related Wikipedia articles in Section E.13.6.

> **Sources**
>
> Background from the Wikipedia article Human Cell Atlas (CC BY-SA 4.0, retrieved 5 October 2026), and from the web pages of the CZI Virtual Cells Platform, the Arc Institute Virtual Cell Challenge and SciLifeLab Alpha Cell (retrieved 5 October 2026).

13.8 Whole-cell models and cryo-electron tomography

The AI virtual cell (Section 13.7) is learned from data. An older and complementary approach builds a model of a cell from its known parts: genes, reactions, molecules and their structures. Such models are explicit about mechanism, but need a complete parts list, which exists only for the simplest cells.

13.8.1 Mechanistic whole-cell models

The first **whole-cell model** simulated the complete life cycle of *Mycoplasma genitalium*, a bacterium with 525 genes. It combined 28 sub-models of different cellular processes, such as metabolism, transcription, translation and DNA replication, each described with the mathematics best suited to it, and predicted phenotypes, such as the growth of single-gene deletion strains, from the genotype (Karr et al. 2012).

The synthetic minimal cell JCVI-syn3A, with 493 genes, has been modelled in more detail. A kinetic model in which the geometry of the cell and the positions of its ribosomes were taken from cryo-electron tomograms reproduced the doubling time and the distribution of mRNA half-lives (Thornburg et al. 2022), and a later model simulated the complete cell cycle, including chromosome segregation and division, in space and time (Thornburg et al. 2026). Such models show which behaviours follow from the known parts, and where knowledge is missing.

13.8.2 Structural models of a cell

Going further, the physical structure of an entire cell can be modelled. Structural models of a whole *Mycoplasma* cell have been built by packing experimental and predicted protein structures, the genome and membranes into a cell volume (Maritan et al. 2022), and work towards simulating an entire minimal cell with molecular dynamics (Section 9.11), at a coarse-grained level, is under way (Stevens et al. 2023). All-atom simulations of a small part of a bacterial cytoplasm showed that crowding and weak interactions between proteins affect their stability and diffusion (Yu et al. 2016). Predicted structures and complexes (Chapter 9 and Chapter 11) provide many of the parts for such models.

13.8.3 Cryo-electron tomography and visual proteomics

Cryo-electron tomography (cryo-ET) images frozen cells, or thin slices of them cut with a focused ion beam, from many angles, and reconstructs a three-dimensional image of the cell interior at close to molecular resolution (Young and Villa 2023; Beck and Baumeister 2016). Unlike single-particle cryo-EM (Section 9.4), it shows complexes in their native context. The idea of **visual proteomics** is to identify and locate every large complex in such tomograms, so that the composition and organisation of a cell can be read from an image (Nickell et al. 2006). In the small bacterium *Mycoplasma pneumoniae*, cryo-ET combined with cross-linking mass spectrometry revealed how RNA polymerase and the ribosome are coupled (O'Reilly et al. 2020), and averaging many copies of the ribosome in cells reached 3.5 Å resolution (Tegunov et al. 2021).

Finding particles in noisy tomograms is an image-analysis problem of the kind described in Section 13.3. Deep-learning methods segment membranes and locate complexes, either trained for specific targets (Teresa-Trueba et al. 2023) or with embeddings that generalise to new ones (Rice et al. 2023), and dedicated tools analyse membranes and membrane proteins (Lamm et al. 2026). The CZ Imaging Institute's **CryoET Data Portal** provides standardised, annotated tomograms for training (Ermel et al. 2024), and a Kaggle competition on particle picking showed how open challenges can improve the methods (Peck et al. 2026). Cryo-ET thus connects the two directions of this chapter: it provides images of whole cells at molecular detail, and the geometry needed by mechanistic and structural models.

See also: further explanation in Section D.13.7; related Wikipedia articles in Section E.13.7.

> **Sources**
>
> Background from the Wikipedia articles Cryogenic electron tomography, Cellular model and Mycoplasma genitalium (CC BY-SA 4.0, retrieved 5 October 2026).

13.9 Key concepts

A checklist for revision.

- **Images as data:** pixels, channels, z-stacks and time series; intensity is not a molecule count; transmitted light, fluorescence, electron microscopy and histology; whole-slide images
- **Classical analysis:** preprocessing, thresholding and watershed segmentation, feature extraction; Fiji, CellProfiler, ilastik, QuPath, napari
- **Data and standards:** metadata (REMBI), OME-Zarr; BioImage Archive, Image Data Resource, BBBC; the Cell Tracking Challenge
- **Tasks:** image and cell classification; semantic vs. instance segmentation; detection and tracking; restoration and label-free prediction

- **Architectures:** CNNs (AlexNet, ResNet); U-Net and skip connections; StarDist (star-convex polygons) and Cellpose (flow fields); vision transformers
- **Self-supervision and foundation models:** DINO and DINOv2, masked autoencoders; Segment Anything and its microscopy versions; why self-supervision matters when labels are scarce
- **Pitfalls:** batch effects between plates, microscopes and hospitals; splitting by experiment; disagreement between annotators; hallucinated structures in predicted images; site signatures in pathology
- **Image-based profiling:** Cell Painting (six dyes, five channels, eight components); single-cell features, aggregation and normalisation to controls; batch correction; profile similarity and mechanism of action; JUMP and RxRx3
- **Protein localisation:** the HPA subcellular atlas; multi-localising proteins and cell-to-cell variation; the HPA Kaggle competitions, multi-label classification and macro-F1; weak supervision; OpenCell, endogenous tagging and cytoself; SubCell; combining image and sequence embeddings
- **Tissues:** CODEX, imaging mass cytometry and MIBI; imaging-based spatial transcriptomics; segmentation, cell-type assignment and neighbourhood analysis; multiple-instance learning on whole slides; pathology foundation models (UNI, CONCH, Virchow, Prov-GigaPath); predicting expression from histology
- **Atlases:** the Human Cell Atlas, HuBMAP, the Allen Institute for Cell Science; what is a cell type?
- **AI virtual cells:** definition; perturbation prediction (GEARS, State); the Virtual Cell Challenge; CZI (Virtual Cells Platform, Billion Cells Project, TranscriptFormer) and SciLifeLab's Alpha Cell; why simple baselines are often as good as deep models
- **Mechanistic models:** the *M. genitalium* whole-cell model; the JCVI-syn3A minimal cell in 4D; structural and molecular-dynamics models of whole cells; learned vs. mechanistic models
- **Cryo-electron tomography:** in situ structural biology; focused-ion-beam milling; subtomogram averaging; visual proteomics; deep learning for particle picking and membrane segmentation; the CryoET Data Portal

13.10 Recommended reading

These papers are recommended for reading in full:

- Chamier et al. (2019) — what a biologist should know about deep learning for microscopy, including its pitfalls.
- Stringer et al. (2020) — Cellpose; a generalist deep-learning method for cell segmentation.
- Ouyang et al. (2019) — image classification of protein localisation as a community challenge, with lessons for evaluation.
- Chandrasekaran et al. (2024) — Cell Painting and image-based profiling of matched chemical and genetic perturbations (JUMP).
- Bunne et al. (2024) — the AI virtual cell; what it is meant to be and how it could be built.
- Ahlmann-Eltze et al. (2025) — a critical test; deep-learning models of perturbation effects against simple baselines.
- Thornburg et al. (2026) — a mechanistic, spatial model of the complete cell cycle of a minimal cell.

Titles, abstracts, links and the full texts are in Section A.13; a longer list of further papers is in Section B.13.

13.11 Background reading (open access)

- **Microscopy (Wikipedia)** — Microscopy is the technical field of using microscopes to view subjects too small to be seen with the naked eye. There are three well-known branches of

microscopy: optical, electron, and scanning probe microscopy, along with the emerging field of X-ray microscopy.
https://en.wikipedia.org/wiki/Microscopy

- **Image segmentation (Wikipedia)** — In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple image segments, also known as image regions or image objects. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images.
https://en.wikipedia.org/wiki/Image_segmentation
- **Convolutional neural network (Wikipedia)** — A convolutional neural network (CNN) is a type of feedforward neural network that learns features via filter optimization. This type of deep learning network has been applied to process and make predictions from many different types of data including text, images and audio.
https://en.wikipedia.org/wiki/Convolutional_neural_network
- **U-Net (Wikipedia)** — U-Net is a convolutional neural network that was developed for image segmentation. The network is based on a fully convolutional neural network whose architecture was modified and extended to work with fewer training images and to yield more precise segmentation. Segmentation of a 512 × 512 image takes less than a second on a modern (2015) GPU using the U-Net architecture.
https://en.wikipedia.org/wiki/U-Net
- **Vision transformer (Wikipedia)** — A vision transformer (ViT) is a transformer designed for computer vision. A ViT decomposes an input image into a series of patches, serializes each patch into a vector, and maps it to a smaller dimension with a single matrix multiplication. These vector embeddings are then processed by a transformer encoder as if they were token embeddings.
https://en.wikipedia.org/wiki/Vision_transformer
- **High-content screening (Wikipedia)** — High-content screening (HCS), also known as high-content analysis (HCA) or cellomics, is a method that is used in biological research and drug discovery to identify substances such as small molecules, peptides, or RNAi that alter the phenotype of a cell in a desired manner.
https://en.wikipedia.org/wiki/High-content_screening
- **Human Protein Atlas (Wikipedia)** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.
https://en.wikipedia.org/wiki/Human_Protein_Atlas
- **Digital pathology (Wikipedia)** — Digital pathology is a sub-field of pathology that focuses on managing and analyzing information generated from digitized specimen slides. It utilizes computer-based technology and virtual microscopy to view, manage, share, and analyze digital slides on computer monitors.
https://en.wikipedia.org/wiki/Digital_pathology
- **Spatial transcriptomics (Wikipedia)** — Spatial transcriptomics, or spatially resolved transcriptomics, is a method that captures positional context of transcriptional activity within intact tissue. The historical precursor to spatial transcriptomics is in situ hybridization, where the modernized omics terminology refers to the measurement of all the mRNA in a cell rather than select RNA targets. It comprises an important part of spatial biology.
https://en.wikipedia.org/wiki/Spatial_transcriptomics
- **Human Cell Atlas (Wikipedia)** — The Human Cell Atlas is a global project to describe all cell types in the human body, to understand human health and for diagnosing, monitoring, and treating disease.
https://en.wikipedia.org/wiki/Human_Cell_Atlas
- **Cellular model (Wikipedia)** — A cellular model or virtual cell is a computational model of

aspects of a biological cell, for the purposes of in silico research.
https://en.wikipedia.org/wiki/Virtual_cell

- **Cryogenic electron tomography (Wikipedia)** — Cryogenic electron tomography (cryoET) is an imaging technique used to reconstruct high-resolution (~1–4 nm) three-dimensional volumes of samples, often biological macromolecules and cells.
https://en.wikipedia.org/wiki/Cryogenic_electron_tomography

Descriptions of Wikipedia articles are taken from the introductions of the articles (CC BY-SA 4.0).

13.12 Student recommendations

Recommended by students preparing for the exam: papers, reviews, online lectures, tutorials and short explanations. To suggest something, use the suggestion form (SU login).

No recommendations yet.

13.13 Review problems

The format of these problems follows Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0); the problems were written for this book, since that book has no chapter on image analysis. In the HTML version, click a box to show the extensions, solution outline and common errors.

13.13.1 Intensity is not abundance

A fluorescence screen reports that the mean nuclear intensity of a tagged protein is twice as high in treated as in untreated cells. What does this show about the amount of the protein in the nucleus, and what would you check before concluding that the treatment increases it?

Extensions, solution outline and common errors

Extensions:

- Which steps of acquisition and analysis affect the measured intensity?
- How do segmentation errors change a per-nucleus mean?
- What metadata are needed for the result to be reusable?

Solution outline:

1. Each pixel value depends on the label, microscope, exposure, detector and background, not only on the number of molecules; images are comparable only if acquired in the same way.
2. Check that treated and untreated wells were imaged with the same settings, on the same plates or with plate controls, and corrected for background and uneven illumination.
3. Segmentation determines which pixels are counted: if the treatment changes nuclear size or shape, the segmentation may include cytoplasm or miss parts of the nucleus, and a mean can change while the total does not (or the reverse).
4. A change in localisation (from cytoplasm to nucleus) is not the same as a change in total amount; measure both compartments.
5. Use replicate wells and plates as the units, not individual cells, and confirm with an independent method such as western blotting or mass spectrometry.

6. Report acquisition and analysis metadata following REMBI, and deposit the images, for example in the BioImage Archive.

Common errors:

- Treating intensity as a molecule count.
- Counting thousands of cells as independent replicates.
- Ignoring that the treatment can change cell shape and so the segmentation.

Connections: Section 13.2 (images as data, classical analysis, metadata); 4 (non-independent examples).

13.13.2 Choosing a segmentation method

A laboratory needs to segment densely packed, touching nuclei and irregularly shaped cells in a new kind of tissue image. Compare thresholding with watershed, StarDist and Cellpose, and describe how you would evaluate the result.

Extensions, solution outline and common errors

Extensions:

- What is the difference between semantic and instance segmentation?
- Why do U-Net skip connections help?
- How can a general model be adapted to new data with few annotations?

Solution outline:

1. Separating touching objects is an instance-segmentation problem; semantic segmentation (nucleus vs. background) is not enough.
2. Thresholding with a watershed transform is fast and transparent but sensitive to uneven staining and tends to over- or under-split.
3. StarDist predicts star-convex polygons, which suits roughly round nuclei but not irregular cell shapes.
4. Cellpose predicts flow fields towards cell centres and works for many shapes without retraining; Cellpose 2.0 can be adapted with a few corrected images (human in the loop).
5. Evaluate on images annotated independently, ideally by more than one person, from experiments not used for training, with object-level measures (for example matched objects at an overlap threshold) rather than pixel accuracy alone.
6. Check that downstream measurements, such as cell counts or marker intensities, are robust to the choice of method.

Common errors:

- Evaluating on tiles from the same images used for training.
- Reporting pixel accuracy, which is dominated by background.
- Assuming that a general model works on a new tissue without checking.

Connections: Section 13.3 (U-Net, StarDist, Cellpose); 4 (evaluation and test sets).

13.13.3 Self-supervised models and foundation models for images

Explain how self-supervised learning can produce useful features for microscopy images without labels, and how you would test whether a pre-trained image model is better than CellProfiler features for classifying protein localisation.

Extensions, solution outline and common errors

Extensions:

- Compare DINO and masked autoencoders.
- Why have self-supervised methods become important for vision transformers?
- What is a fair comparison?

Solution outline:

1. Vision transformers treat image patches as tokens and need large training sets; labels for microscopy are scarce, which makes self-supervision attractive.
2. In DINO, a student network learns to match a teacher network on different views of the same image; masked autoencoders hide most patches and reconstruct them, like masked language modelling.
3. Features are evaluated by probing: train a simple classifier, for example logistic regression, on frozen embeddings, and compare with the same classifier on CellProfiler features.
4. Split by protein, cell line or experiment rather than by image or cell, so that the same protein does not occur in both training and test sets.
5. Use measures suited to multi-label, imbalanced classes (for example macro-F1 or per-class precision–recall).
6. Check that the embeddings do not mainly encode plate, batch or microscope.

Common errors:

- Splitting by cell, so that images of the same protein are on both sides.
- Comparing a tuned deep model with an untuned baseline.
- Assuming that a model pre-trained on natural images, such as SAM, works on microscopy without fine-tuning or testing.

Connections: Section 13.3 (vision transformers and self-supervision); Section 13.5 (HPA and image-based localisation); Section 12.6 (evaluating foundation models).

13.13.4 Image-based profiling

Describe how a Cell Painting experiment is turned into profiles, and how such profiles can be used to suggest the mechanism of action of an uncharacterised compound.

Extensions, solution outline and common errors

Extensions:

- What do the six dyes label?
- Why normalise to negative controls on the same plate?
- How would you test whether a compound has any effect at all?

Solution outline:

1. Six dyes, imaged in five channels, label DNA, RNA and nucleoli, the endoplasmic reticulum, actin, Golgi and plasma membrane, and mitochondria.
2. Cells are segmented and hundreds to thousands of size, shape, intensity and texture features (or deep embeddings) are measured per cell.
3. Features are aggregated per well, normalised against negative controls on the same plate, reduced to informative, uncorrelated features, and corrected for batch effects between plates and sites.
4. Test whether a compound's profile differs from controls more than replicate controls differ from each other.
5. Compare its profile with profiles of compounds of known mechanism, or of genetic perturbations, for example by cosine similarity; similar profiles suggest a shared mechanism, a hypothesis to be tested.
6. Evaluate such matching with held-out plates and replicate consistency, as in the CPJUMP1 benchmark.

Common errors:

- Interpreting profile similarity as proof of a shared target.
- Ignoring plate and batch effects, so that compounds on the same plate look similar.
- Using cells rather than wells or plates as replicates.

Connections: Section 13.4 (Cell Painting, JUMP, RxRx3); Section 7.4 (normalisation, batch effects, clustering).

13.13.5 Site signatures in computational pathology

A model trained on whole-slide images from The Cancer Genome Atlas predicts patient survival with high accuracy. Explain how site signatures could inflate this result and how you would design a fair evaluation.

Extensions, solution outline and common errors

Extensions:

- How is a whole-slide image turned into a single prediction?
- Why does colour normalisation not solve the problem?
- What role can pathology foundation models play?

Solution outline:

1. Slides are split into tiles, each tile is embedded, and the embeddings are aggregated, often with attention, in multiple-instance learning.
2. Staining, scanning and preparation differ between hospitals, and the submitting site can be recognised from the image even after colour normalisation.
3. If survival or patient mix differs between sites, a model can predict outcome by recognising the site; with slides from the same sites in training and test sets, accuracy is inflated.
4. Split by site, keep all slides from a patient together, and test on an external cohort.
5. Check whether the site can be predicted from the embeddings and whether performance holds within sites.

6. Foundation models such as UNI or Virchow give general features, but their embeddings can also encode site, and their pre-training data may overlap with the test slides.

Common errors:

- Splitting by tile or slide rather than by patient and site.
- Assuming colour normalisation removes all site effects.
- Interpreting attention maps as the cause of the prediction.

Connections: Section 13.6 (computational pathology); 4 (data leakage and external validation); Section 12.4 (interpretability).

13.13.6 Spatial analysis of multiplexed images

A study of multiplexed tissue images (for example imaging mass cytometry) reports that T cells are found next to tumour cells more often than expected in patients who respond to therapy. Describe the analysis, and what is needed before the result is trusted.

Extensions, solution outline and common errors

Extensions:

- What are the steps from image to cell types?
- What is a suitable null model for neighbourhood enrichment?
- What is the independent unit?

Solution outline:

1. Segment the cells, quantify each marker per cell, and assign cell types by clustering or by a classifier such as CellSighter.
2. Segmentation errors cause spill-over of markers between neighbouring cells, which can create apparent mixed cell types and spurious neighbours.
3. Neighbourhood enrichment is tested against permutations of cell-type labels within each image, which keep the positions and density of cells.
4. Tissue architecture (for example tumour borders) and the area sampled can cause enrichment by themselves.
5. The patient is the unit: many cells or regions from one patient are not independent replicates, and the comparison between responders and non-responders must be made between patients, with an independent cohort.
6. Spatial association is not causation; a role of T-cell contact would need perturbation or longitudinal data.

Common errors:

- Treating each cell or image region as an independent sample.
- Using a null model that ignores tissue structure.
- Ignoring marker spill-over from segmentation errors.

Connections: Section 13.6 (multiplexed imaging and spatial omics); Section 7.6 (spatial transcriptomics and cell types); Section 12.8 (prediction and causation).

13.13.7 How good is a virtual cell?

A company claims that its AI virtual cell model predicts the transcriptional response of human cells to genetic perturbations that were never measured. What evidence would you require?

Extensions, solution outline and common errors

Extensions:

- What baselines should be used?
- How should the test perturbations be chosen?
- Why did the first Virtual Cell Challenge find no single good measure?

Solution outline:

1. Define what is predicted (expression changes, cell-type shifts, combinations) and in which cell types.
2. Hold out perturbations, and ideally cell types or experiments, that were not in training; combinations should not have their single perturbations in the test set if that is the claim.
3. Compare with simple baselines, such as predicting the mean response or a linear model, which in independent evaluations have matched deep models.
4. Use several measures (correlation of changes, error on differentially expressed genes, ranking of perturbations), since a model can look good on one and poor on another.
5. Prefer blind benchmarks, such as the Virtual Cell Challenge, in which predictions are made before the data are released.
6. Confirm selected predictions experimentally.

Common errors:

- Accepting performance on perturbations similar to the training data as generalisation.
- Omitting simple baselines.
- Reporting only the most favourable measure.

Connections: Section 13.7 (virtual cells and their evaluation); Section 12.6 (foundation models and baselines); Section 9.2 (blind benchmarks).

13.13.8 Mechanistic and learned models of cells

Compare a mechanistic whole-cell model, such as that of *Mycoplasma genitalium* or JCVI-syn3A, with a learned AI virtual cell. What can each do, and how can cryo-electron tomography contribute to both?

Extensions, solution outline and common errors

Extensions:

- Why have mechanistic models been built only for minimal cells?
- What is visual proteomics?
- What are the image-analysis problems of cryo-ET?

Solution outline:

1. Mechanistic models are built from known genes, reactions and structures, combine sub-models of processes such as metabolism and replication, and predict phenotypes, such as

the growth of deletion strains, from genotype; they need a nearly complete parts list and many parameters.
2. Learned models are trained on atlases, perturbation screens and images; they can cover human cells but are only as good as their training data and must be tested against simple baselines.
3. Mechanistic models show which behaviours follow from known parts and where knowledge is missing; learned models may predict without explaining.
4. Cryo-ET gives three-dimensional images of the cell interior near molecular resolution; visual proteomics aims to identify and locate every large complex.
5. Finding particles in noisy tomograms is a deep-learning problem; annotated data such as the CryoET Data Portal and open challenges help.
6. The geometry from cryo-ET (cell shape, ribosome positions) was used in the JCVI-syn3A model, and structural maps can provide training data and parts for both kinds of model.

Common errors:

- Assuming either kind of model is complete or correct because it reproduces some observations.
- Confusing cryo-ET with single-particle cryo-EM.

Connections: Section 13.8 (whole-cell models and cryo-ET); Section 13.7 (AI virtual cells); Section 9.4 (cryo-EM); Section 11.3 (systems biology and metabolic models).

References

"A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium." 2014. *Nature Biotechnology* 32 (9): 903–14. https://doi.org/10.1038/nbt.2957.

"A Haplotype Map of the Human Genome." 2005. *Nature* 437 (7063): 1299–320. https://doi.org/10.1038/nature04226.

"A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome." 1994. *Molecular Biology and Evolution*, ahead of print. https://doi.org/10.1093/oxfordjournals.molbev.a040152.

Abascal, Federico, David Juan, Irwin Jungreis, et al. 2018. "Loose Ends: Almost One in Five Human Genes Still Have Unresolved Coding Status." *Nucleic Acids Research* 46 (14): 7070–84. https://doi.org/10.1093/nar/gky587.

Abramson, Josh, Jonas Adler, Jack Dunger, et al. 2024. "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3." *Nature* 630 (8016): 493–500. https://doi.org/10.1038/s41586-024-07487-w.

Abriata, Luciano A., and Matteo Dal Peraro. 2025. "Practical Outcomes from CASP16 for Users in Need of Biomolecular Structure Prediction." *Proteins: Structure, Function, and Bioinformatics* 94 (1): 435–46. https://doi.org/10.1002/prot.70078.

Adames, Neil R., Jenna E. Gallegos, and Jean Peccoud. 2018. "Yeast Genetic Interaction Screens in the Age of CRISPR/Cas." *Current Genetics* 65 (2): 307–27. https://doi.org/10.1007/s00294-018-0887-8.

Adams, David, Richard Baldock, Shoumo Bhattacharya, et al. 2013. "Bloomsbury Report on Mouse Embryo Phenotyping: Recommendations from the IMPC Workshop on Embryonic Lethal Screening." *Disease Models & Mechanisms*, ahead of print. https://doi.org/10.1242/dmm.011833.

Adams, Mark D., Susan E. Celniker, Robert A. Holt, et al. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science* 287 (5461): 2185–95. https://doi.org/10.1126/science.287.5461.2185.

Addona, Terri A, Susan E Abbatiello, Birgit Schilling, et al. 2009. "Multi-Site Assessment of the Precision and Reproducibility of Multiple Reaction Monitoring–Based Measurements of Proteins in Plasma." *Nature Biotechnology* 27 (7): 633–41. https://doi.org/10.1038/nbt.1546.

Adduri, Abhinav K., Dhruv Gautam, Beatrice Bevilacqua, et al. 2026. "Predicting Cellular Responses to Perturbation Across Diverse Contexts with State." *Cell* 189 (19): 5914–5931.e20. https://doi.org/10.1016/j.cell.2026.07.052.

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, et al. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4): 248–49. https://doi.org/10.1038/nmeth0410-248.

Aebersold, Ruedi, and Matthias Mann. 2003. "Mass Spectrometry-Based Proteomics." *Nature* 422 (6928): 198–207. https://doi.org/10.1038/nature01511.

Aebersold, Ruedi, and Matthias Mann. 2016. "Mass-Spectrometric Exploration of Proteome Structure and Function." *Nature* 537 (7620): 347–55. https://doi.org/10.1038/nature19949.

Aguirre, Andrew J., Robin M. Meyers, Barbara A. Weir, et al. 2016. "Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting." *Cancer Discovery* 6 (8): 914–29. https://doi.org/10.1158/2159-8290.cd-16-0154.

Ahdritz, Gustaf, Nazim Bouatta, Christina Floristean, et al. 2024. "OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization." *Nature Methods* 21 (8): 1514–24. https://doi.org/10.1038/s41592-024-02272-z.

Ahlmann-Eltze, Constantin, Wolfgang Huber, and Simon Anders. 2025. "Deep-Learning-Based Gene Perturbation Effect Prediction Does Not yet Outperform Simple Linear Baselines." *Nature Methods* 22 (8): 1657–61. https://doi.org/10.1038/s41592-025-02772-6.

Ahlquist, K D, Mayra M Bañuelos, Alyssa Funk, et al. 2021. "Our Tangled Family Tree: New Genomic Methods Offer Insight into the Legacy of Archaic Admixture." *Genome Biology and Evolution* 13 (7). https://doi.org/10.1093/gbe/evab115.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. *Molecular Biology of the Cell.* 4th ed. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK21054/.

Alderson, T. Reid, Iva Pritišanac, Đesika Kolarić, Alan M. Moses, and Julie D. Forman-Kay. 2023. "Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by AlphaFold2." *Proceedings of the National Academy of Sciences* 120 (44). https://doi.org/10.1073/pnas.2304302120.

Alexa, Adrian, Jörg Rahnenführer, and Thomas Lengauer. 2006. "Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure." *Bioinformatics* 22 (13): 1600–1607. https://doi.org/10.1093/bioinformatics/btl140.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, et al. 2013. "Signatures of Mutational Processes in Human Cancer." *Nature* 500 (7463): 415–21. https://doi.org/10.1038/nature12477.

Alexeyenko, Andrey, Woojoo Lee, Maria Pernemalm, et al. 2012. "Network Enrichment Analysis: Extension of Gene-Set Enrichment Analysis to Gene Networks." *BMC Bioinformatics* 13 (1). https://doi.org/10.1186/1471-2105-13-226.

Alexeyenko, Andrey, and Erik L. L. Sonnhammer. 2009. "Global Networks of Functional Coupling in Eukaryotes from Comprehensive Data Integration." *Genome Research* 19 (6): 1107–16. https://doi.org/10.1101/gr.087528.108.

Alfaro, Javier Antonio, Peggy Bohländer, Mingjie Dai, et al. 2021. "The Emerging Landscape of Single-Molecule Protein Sequencing Technologies." *Nature Methods* 18 (6): 604–17. https://doi.org/10.1038/s41592-021-01143-1.

Alford, Rebecca F., Andrew Leaver-Fay, Jeliazko R. Jeliazkov, et al. 2017. "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design." *Journal of Chemical Theory and Computation* 13 (6): 3031–48. https://doi.org/10.1021/acs.jctc.7b00125.

Alipanahi, Babak, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." *Nature Biotechnology* 33 (8): 831–38. https://doi.org/10.1038/nbt.3300.

Alley, Ethan C., Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. 2019. "Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning." *Nature Methods* 16 (12): 1315–22. https://doi.org/10.1038/s41592-019-0598-1.

Almagro Armenteros, Jose Juan, Marco Salvatore, Olof Emanuelsson, et al. 2019. "Detecting Sequence Signals in Targeting Peptides Using Deep Learning." *Life Science Alliance* 2 (5): e201900429. https://doi.org/10.26508/lsa.201900429.

Almagro Armenteros, José Juan, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. 2017. "DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning." *Bioinformatics* 33 (21): 3387–95. https://doi.org/10.1093/bioinformatics/btx431.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, et al. 2019. "A New Genomic Blueprint of the Human Gut Microbiota." *Nature* 568 (7753): 499–504. https://doi.org/10.1038/s41586-019-0965-1.

Alon, Uri. 2007. "Network Motifs: Theory and Experimental Approaches." *Nature Reviews Genetics* 8 (6): 450–61. https://doi.org/10.1038/nrg2102.

Alon, U., M. G. Surette, N. Barkai, and S. Leibler. 1999. "Robustness in Bacterial Chemotaxis." *Nature* 397 (6715): 168–71. https://doi.org/10.1038/16483.

Alser, Mohammed, Jeremy Rotman, Dhrithi Deshpande, et al. 2021. "Technology Dictates Algorithms: Recent Developments in Read Alignment." *Genome Biology* 22 (1). https://doi.org/10.1186/s13059-021-02443-7.

Altemose, Nicolas, Glennis A. Logsdon, Andrey V. Bzikadze, et al. 2022. "Complete Genomic and Epigenetic Maps of Human Centromeres." *Science* 376 (6588). https://doi.org/10.1126/science.abl4178.

Altenhoff, Adrian M, Brigitte Boeckmann, Salvador Capella-Gutierrez, et al. 2016. "Standardized Benchmarking in the Quest for Orthologs." *Nature Methods* 13 (5): 425–30. https://doi.org/10.1038/nmeth.3830.

Altenhoff, Adrian M., Manuel Gil, Gaston H. Gonnet, and Christophe Dessimoz. 2013. "Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs." *PLoS ONE* 8 (1): e53786. https://doi.org/10.1371/journal.pone.0053786.

Altenhoff, Adrian M., Natasha M. Glover, and Christophe Dessimoz. 2019. "Inferring Orthology and Paralogy." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_5.

Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. 2012. "Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function Than Paralogs." *PLoS Computational Biology* 8 (5): e1002514. https://doi.org/10.1371/jo

urnal.pcbi.1002514.

Altenhoff, Adrian M, Clément-Marie Train, Kimberly J Gilbert, et al. 2020. "OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More." *Nucleic Acids Research* 49 (D1): D373–79. https://doi.org/10.1093/nar/gkaa1007.

Altschul, Stephen F. 1991. "Amino Acid Substitution Matrices from an Information Theoretic Perspective." *Journal of Molecular Biology* 219 (3): 555–65. https://doi.org/10.1016/0022-2836(91)90193-a.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/s0022-2836(05)80360-2.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, et al. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–402. https://doi.org/10.1093/nar/25.17.3389.

Amaral, Paulo, Silvia Carbonell-Sala, Francisco M. De La Vega, et al. 2023. "The Status of the Human Gene Catalogue." *Nature* 622 (7981): 41–47. https://doi.org/10.1038/s41586-023-06490-x.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-1935-5.

Amberger, Joanna S, Carol A Bocchini, Alan F Scott, and Ada Hamosh. 2018. "OMIM.org: Leveraging Knowledge Across Phenotype–Gene Relationships." *Nucleic Acids Research* 47 (D1): D1038–43. https://doi.org/10.1093/nar/gky1151.

Ambroise, Christophe, and Geoffrey J. McLachlan. 2002. "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data." *Proceedings of the National Academy of Sciences* 99 (10): 6562–66. https://doi.org/10.1073/pnas.102102699.

Amitay, Yael, Yuval Bussi, Ben Feinstein, Shai Bagon, Idan Milo, and Leeat Keren. 2023. "CellSighter: A Neural Network to Classify Cells in Highly Multiplexed Images." *Nature Communications* 14 (1). https://doi.org/10.1038/s41467-023-40066-7.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10). https://doi.org/10.1186/gb-2010-11-10-r106.

Anderson, Douglas M., Kelly M. Anderson, Chi-Lun Chang, et al. 2015. "A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance." *Cell* 160 (4): 595–606. https://doi.org/10.1016/j.cell.2015.01.009.

Anderson, S., A. T. Bankier, B. G. Barrell, et al. 1981. "Sequence and Organization of the Human Mitochondrial Genome." *Nature* 290 (5806): 457–65. https://doi.org/10.1038/290457a0.

Andreeva, Antonina, Eugene Kulesha, Julian Gough, and Alexey G Murzin. 2019. "The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures." *Nucleic Acids Research* 48 (D1): D376–82. https://doi.org/10.1093/nar/gkz1064.

Anfinsen, Christian B. 1973. "Principles That Govern the Folding of Protein Chains." *Science* 181 (4096):

223–30. https://doi.org/10.1126/science.181.4096.223.

Angelis, Jesse, Eva Ayla Schröder, Zixuan Xiao, Wassim Gabriel, and Mathias Wilhelm. 2025. "Peptide Property Prediction for Mass Spectrometry Using AI: An Introduction to State of the Art Models." *PROTEOMICS* 25 (9-10). https://doi.org/10.1002/pmic.202400398.

Angelo, Michael, Sean C Bendall, Rachel Finck, et al. 2014. "Multiplexed Ion Beam Imaging of Human Breast Tumors." *Nature Medicine* 20 (4): 436–42. https://doi.org/10.1038/nm.3488.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. 2016. "Deep Learning for Computational Biology." *Molecular Systems Biology* 12 (7). https://doi.org/10.15252/msb.20156651.

Anishchenko, Ivan, Samuel J. Pellock, Tamuka M. Chidyausiku, et al. 2021. "De Novo Protein Design by Deep Network Hallucination." *Nature* 600 (7889): 547–52. https://doi.org/10.1038/s41586-021-04184-w.

Anisimova, Maria, and Olivier Gascuel. 2006. "Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative." *Systematic Biology* 55 (4): 539–52. https://doi.org/10.1080/10635150600755453.

Anisimova, Maria, Rasmus Nielsen, and Ziheng Yang. 2003. "Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites." *Genetics* 164 (3): 1229–36. https://doi.org/10.1093/genetics/164.3.1229.

Anson, Eric L., and Eugene W. Myers. 1999. "Algorithms for Whole Genome Shotgun Sequencing." *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB '99)*, 1–9.

Apic, Gordana, Julian Gough, and Sarah A Teichmann. 2001. "Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes." *Journal of Molecular Biology* 310 (2): 311–25. https://doi.org/10.1006/jmbi.2001.4776.

Archit, Anwai, Luca Freckmann, Sushmita Nair, et al. 2025. "Segment Anything for Microscopy." *Nature Methods* 22 (3): 579–91. https://doi.org/10.1038/s41592-024-02580-4.

Argelaguet, Ricard, Damien Arnol, Danila Bredikhin, et al. 2020. "MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02015-1.

Argelaguet, Ricard, Britta Velten, Damien Arnol, et al. 2018. "Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-Omics Data Sets." *Molecular Systems Biology* 14 (6). https://doi.org/10.15252/msb.20178124.

Aris-Brosou, Stéphane, and Nicolas Rodrigue. 2019. "A Not-so-Long Introduction to Computational Molecular Evolution." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_3.

Arita, Masanori, Ilene Karsch-Mizrachi, and Guy Cochrane. 2020. "The International Nucleotide Sequence Database Collaboration." *Nucleic Acids Research* 49 (D1): D121–24. https://doi.org/10.1093/nar/gkaa967.

Armstrong, Joel, Ian T. Fiddes, Mark Diekhans, and Benedict Paten. 2019. "Whole-Genome Alignment and Comparative Annotation." *Annual Review of Animal Biosciences* 7 (1): 41–64. https://doi.org/10.1146/annurev-animal-020518-115005.

Armstrong, Joel, Glenn Hickey, Mark Diekhans, et al. 2020. "Progressive Cactus Is a Multiple-Genome Aligner for the Thousand-Genome Era." *Nature* 587 (7833): 246–51. https://doi.org/10.1038/s41586-020-2871-y.

Asgari, Ehsaneddin, and Mohammad R. K. Mofrad. 2015. "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics." *PLOS ONE* 10 (11): e0141287. https://doi.org/10.1371/journal.pone.0141287.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Ashkenazy, Haim, Shiran Abadi, Eric Martz, et al. 2016. "ConSurf 2016: An Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules." *Nucleic Acids Research* 44 (W1): W344–50. https://doi.org/10.1093/nar/gkw408.

Assarsson, Erika, Martin Lundberg, Göran Holmquist, et al. 2014. "Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability." *PLoS ONE* 9 (4): e95192. https://doi.org/10.1371/journal.pone.0095192.

Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. 2001. "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model." *Biophysical Journal* 80 (1): 505–15. https://doi.org/10.1016/s0006-3495(01)76033-x.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, et al. 2021. "Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions." *Nature Methods* 18 (10): 1196–203. https://doi.org/10.1038/s41592-021-01252-x.

Avsec, Žiga, Natasha Latysheva, Jun Cheng, et al. 2026. "Advancing Regulatory Variant Effect Prediction with AlphaGenome." *Nature* 649 (8099): 1206–18. https://doi.org/10.1038/s41586-025-10014-0.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, et al. 2021. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network." *Science* 373 (6557): 871–76. https://doi.org/10.1126/science.abj8754.

Baek, Seungbyn, Kyungwoo Song, and Insuk Lee. 2025. "Single-Cell Foundation Models: Bringing Artificial Intelligence into Cell Biology." *Experimental & Molecular Medicine* 57 (10): 2169–81. https://doi.org/10.1038/s12276-025-01547-5.

Bagos, P. G., T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas. 2004. "PRED-TMBB: A Web Server for Predicting the Topology of -Barrel Outer Membrane Proteins." *Nucleic Acids Research* 32 (Web Server): W400–404. https://doi.org/10.1093/nar/gkh417.

Bahar, Ivet, Ali Rana Atilgan, and Burak Erman. 1997. "Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential." *Folding and Design* 2 (3): 173–81. https://doi.org/10.1016/s1359-0278(97)00024-2.

Bailey, Jeffrey A., Zhiping Gu, Royden A. Clark, et al. 2002. "Recent Segmental Duplications in the

Human Genome." *Science* 297 (5583): 1003–7. https://doi.org/10.1126/science.1072047.

Bailey, Timothy L., and Charles Elkan. 1994. "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers." *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 28–36.

Bairoch, A. 1991. "Prosite: A Dictionary of Sites and Patterns in Proteins." *Nucleic Acids Research* 19 (suppl): 2241–45. https://doi.org/10.1093/nar/19.suppl.2241.

Baltimore, D. 1971. "Expression of Animal Virus Genomes." *Bacteriological Reviews* 35 (3): 235–41. https://doi.org/10.1128/br.35.3.235-241.1971.

Bamshad, Michael J., Sarah B. Ng, Abigail W. Bigham, et al. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews Genetics* 12 (11): 745–55. https://doi.org/10.1038/nrg3031.

Ban, Nenad, Poul Nissen, Jeffrey Hansen, Peter B. Moore, and Thomas A. Steitz. 2000. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 å Resolution." *Science* 289 (5481): 905–20. https://doi.org/10.1126/science.289.5481.905.

Banani, Salman F., Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. 2017. "Biomolecular Condensates: Organizers of Cellular Biochemistry." *Nature Reviews Molecular Cell Biology* 18 (5): 285–98. https://doi.org/10.1038/nrm.2017.7.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bankhead, Peter. 2022. *Introduction to Bioimage Analysis*. Online book. https://bioimagebook.github.io/.

Bankhead, Peter, Maurice B. Loughrey, José A. Fernández, et al. 2017. "QuPath: Open Source Software for Digital Pathology Image Analysis." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-17204-5.

Bansal, Parit, Anne Morgat, Kristian B Axelsen, et al. 2021. "Rhea, the Reaction Knowledgebase in 2022." *Nucleic Acids Research* 50 (D1): D693–700. https://doi.org/10.1093/nar/gkab1016.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–12. https://doi.org/10.1126/science.286.5439.509.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2010. "Network Medicine: A Network-Based Approach to Human Disease." *Nature Reviews Genetics* 12 (1): 56–68. https://doi.org/10.1038/nrg2918.

Barabási, Albert-László, and Zoltán N. Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5 (2): 101–13. https://doi.org/10.1038/nrg1272.

Barbosa-Morais, Nuno L., Manuel Irimia, Qun Pan, et al. 2012. "The Evolutionary Landscape of Alternative Splicing in Vertebrate Species." *Science* 338 (6114): 1587–93. https://doi.org/10.1126/science.1230612.

Barkai, N., and S. Leibler. 1997. "Robustness in Simple Biochemical Networks." *Nature* 387 (6636): 913–17. https://doi.org/10.1038/43199.

Barrio-Hernandez, Inigo, Jingi Yeo, Jürgen Jänes, et al. 2023. "Clustering Predicted Structures at the Scale of the Known Protein Universe." *Nature* 622 (7983): 637–45. https://doi.org/10.1038/s41586-023-06510-w.

Bartel, David P. 2018. "Metazoan MicroRNAs." *Cell* 173 (1): 20–51. https://doi.org/10.1016/j.cell.2018.03.006.

Basu, Malay Kumar, Liran Carmel, Igor B. Rogozin, and Eugene V. Koonin. 2008. "Evolution of Protein Domain Promiscuity in Eukaryotes." *Genome Research* 18 (3): 449–61. https://doi.org/10.1101/gr.6943508.

Basu, Sankar, and Björn Wallner. 2016. "DockQ: A Quality Measure for Protein-Protein Docking Models." *PLOS ONE* 11 (8): e0161879. https://doi.org/10.1371/journal.pone.0161879.

Bateman, Alex, Maria-Jesus Martin, Sandra Orchard, et al. 2024. "UniProt: The Universal Protein Knowledgebase in 2025." *Nucleic Acids Research* 53 (D1): D609–17. https://doi.org/10.1093/nar/gkae1010.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics* 41 (1): 164–71. https://doi.org/10.1214/aoms/1177697196.

Beausoleil, Sean A, Judit Villén, Scott A Gerber, John Rush, and Steven P Gygi. 2006. "A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization." *Nature Biotechnology* 24 (10): 1285–92. https://doi.org/10.1038/nbt1240.

Becht, Etienne, Leland McInnes, John Healy, et al. 2018. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology* 37 (1): 38–44. https://doi.org/10.1038/nbt.4314.

Beck, Martin, and Wolfgang Baumeister. 2016. "Cryo-Electron Tomography: Can It Reveal the Molecular Sociology of Cells in Atomic Detail?" *Trends in Cell Biology* 26 (11): 825–37. https://doi.org/10.1016/j.tcb.2016.08.006.

Beerenwinkel, Niko, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. 2014. "Cancer Evolution: Mathematical Models and Computational Inference." *Systematic Biology* 64 (1): e1–25. https://doi.org/10.1093/sysbio/syu081.

Bell, Elizabeth A., Patrick Boehnke, T. Mark Harrison, and Wendy L. Mao. 2015. "Potentially Biogenic Carbon Preserved in a 4.1 Billion-Year-Old Zircon." *Proceedings of the National Academy of Sciences* 112 (47): 14518–21. https://doi.org/10.1073/pnas.1517557112.

Bellman, Richard. 1954. "The Theory of Dynamic Programming." *Bulletin of the American Mathematical Society* 60 (6): 503–15. https://doi.org/10.1090/s0002-9904-1954-09848-8.

Ben-Hur, Asa, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. 2008. "Support Vector Machines and Kernels for Computational Biology." *PLoS Computational Biology* 4 (10): e1000173. https://doi.org/10.1371/journal.pcbi.1000173.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59. https://doi.org/10.1038/nature07517.

Bepler, Tristan, and Bonnie Berger. 2021. "Learning the Protein Language: Evolution, Structure, and Function." *Cell Systems* 12 (6): 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

Berg, Stuart, Dominik Kutra, Thorben Kroeger, et al. 2019. "Ilastik: Interactive Machine Learning for (Bio)image Analysis." *Nature Methods* 16 (12): 1226–32. https://doi.org/10.1038/s41592-019-0582-9.

Berget, Susan M., Claire Moore, and Phillip A. Sharp. 1977. "Spliced Segments at the 5′ Terminus of Adenovirus 2 Late mRNA." *Proceedings of the National Academy of Sciences* 74 (8): 3171–75. https://doi.org/10.1073/pnas.74.8.3171.

Berman, Helen M., John Westbrook, Zukang Feng, et al. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42. https://doi.org/10.1093/nar/28.1.235.

Bernett, Judith, David B Blumenthal, and Markus List. 2024. "Cracking the Black Box of Deep Sequence-Based Protein–Protein Interaction Prediction." *Briefings in Bioinformatics* 25 (2). https://doi.org/10.1093/bib/bbae076.

Bertoline, Letícia M. F., Angélica N. Lima, Jose E. Krieger, and Samantha K. Teixeira. 2023. "Before and After AlphaFold2: An Overview of Protein Structure Prediction." *Frontiers in Bioinformatics* 3. https://doi.org/10.3389/fbinf.2023.1120370.

Bertoni, Damian, Maxim Tsenkov, Paulyna Magana, et al. 2025. "AlphaFold Protein Structure Database 2025: A Redesigned Interface and Updated Structural Coverage." *Nucleic Acids Research* 54 (D1): D358–62. https://doi.org/10.1093/nar/gkaf1226.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, and Davide Pisani. 2018. "Integrated Genomic and Fossil Evidence Illuminates Life's Early Evolution and Eukaryote Origin." *Nature Ecology & Evolution* 2 (10): 1556–62. https://doi.org/10.1038/s41559-018-0644-x.

Bileschi, Maxwell L., David Belanger, Drew H. Bryant, et al. 2022. "Using Deep Learning to Annotate the Protein Universe." *Nature Biotechnology* 40 (6): 932–37. https://doi.org/10.1038/s41587-021-01179-w.

Bininda-Emonds, Olaf R. P. 2004. "The Evolution of Supertrees." *Trends in Ecology & Evolution* 19 (6): 315–22. https://doi.org/10.1016/j.tree.2004.03.015.

Birchler, James A, and Hua Yang. 2022. "The Multiple Fates of Gene Duplications: Deletion, Hypofunctionalization, Subfunctionalization, Neofunctionalization, Dosage Balance Constraints, and Neutral Variation." *The Plant Cell* 34 (7): 2466–74. https://doi.org/10.1093/plcell/koac076.

Birney, Ewan. 2021. "The International Human Genome Project." *Human Molecular Genetics* 30 (R2): R161–63. https://doi.org/10.1093/hmg/ddab198.

Bitbol, Anne-Florence, Robert S. Dwyer, Lucy J. Colwell, and Ned S. Wingreen. 2016. "Inferring Interaction Partners from Protein Sequences." *Proceedings of the National Academy of Sciences* 113 (43): 12180–85. https://doi.org/10.1073/pnas.1606762113.

Björklund, Åsa K., Diana Ekman, Sara Light, Johannes Frey-Skött, and Arne Elofsson. 2005. "Domain Rearrangements in Protein Evolution." *Journal of Molecular Biology* 353 (4): 911–23. https://doi.org/10.1016/j.jmb.2005.08.067.

Björklund, Åsa K, Diana Ekman, and Arne Elofsson. 2006. "Expansion of Protein Domain Repeats." *PLoS Computational Biology* 2 (8): e114. https://doi.org/10.1371/journal.pcbi.0020114.

Bjornson, Saelin, Heroen Verbruggen, Nathan S. Upham, and Jacob L. Steenwyk. 2024. "Reticulate Evolution: Detection and Utility in the Phylogenomics Era." *Molecular Phylogenetics and Evolution* 201: 108197. https://doi.org/10.1016/j.ympev.2024.108197.

Blanchette, Mathieu, W. James Kent, Cathy Riemer, et al. 2004. "Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner." *Genome Research* 14 (4): 708–15. https://doi.org/10.1101/gr.1933104.

Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, et al. 2023. "Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4." *Nature Biotechnology* 41 (11): 1633–44. https://doi.org/10.1038/s41587-023-01688-w.

Blattner, Frederick R., Guy Plunkett, Craig A. Bloch, et al. 1997. "The Complete Genome Sequence of Escherichia Coli k-12." *Science* 277 (5331): 1453–62. https://doi.org/10.1126/science.277.5331.1453.

Blobel, G, and B Dobberstein. 1975. "Transfer of Proteins Across Membranes. I. Presence of Proteolytically Processed and Unprocessed Nascent Immunoglobulin Light Chains on Membrane-Bound Ribosomes of Murine Myeloma." *The Journal of Cell Biology* 67 (3): 835–51. https://doi.org/10.1083/jcb.67.3.835.

Blom, Nikolaj, Steen Gammeltoft, and Søren Brunak. 1999. "Sequence and Structure-Based Prediction of Eukaryotic Protein Phosphorylation Sites." *Journal of Molecular Biology* 294 (5): 1351–62. https://doi.org/10.1006/jmbi.1999.3310.

Blom, Nikolaj, Thomas Sicheritz-Pontén, Ramneek Gupta, Steen Gammeltoft, and Søren Brunak. 2004. "Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence." *PROTEOMICS* 4 (6): 1633–49. https://doi.org/10.1002/pmic.200300771.

Blum, Matthias, Antonina Andreeva, Laise Cavalcanti Florentino, et al. 2024. "InterPro: The Protein Sequence Classification Resource in 2025." *Nucleic Acids Research* 53 (D1): D444–56. https://doi.org/10.1093/nar/gkae1082.

Blum, Matthias, Hsin-Yu Chang, Sara Chuguransky, et al. 2020. "The InterPro Protein Families and Domains Database: 20 Years On." *Nucleic Acids Research* 49 (D1): D344–54. https://doi.org/10.1093/nar/gkaa977.

Bobay, Louis-Marie, and Howard Ochman. 2017. "The Evolution of Bacterial Genome Architecture." *Frontiers in Genetics* 8. https://doi.org/10.3389/fgene.2017.00072.

Bock, Christoph, Paul Datlinger, Florence Chardon, et al. 2022. "High-Content CRISPR Screening."

Nature Reviews Methods Primers 2 (1). https://doi.org/10.1038/s43586-021-00093-4.

Bogan, Andrew A, and Kurt S Thorn. 1998. "Anatomy of Hot Spots in Protein Interfaces." *Journal of Molecular Biology* 280 (1): 1–9. https://doi.org/10.1006/jmbi.1998.1843.

Böhm, Hans-Joachim. 1994. "The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure." *Journal of Computer-Aided Molecular Design* 8 (3): 243–56. https://doi.org/10.1007/bf00126743.

Boiarsky, Rebecca, Nalini M. Singh, Alejandro Buendia, Ava P. Amini, Gad Getz, and David Sontag. 2024. "Deeper Evaluation of a Single-Cell Foundation Model." *Nature Machine Intelligence* 6 (12): 1443–46. https://doi.org/10.1038/s42256-024-00949-w.

Boitreaud, Jacques, Jack Dent, Matthew McPartlon, et al. 2024. "Chai-1: Decoding the Molecular Interactions of Life." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2024.10.10.615955.

Bolstad, B. M., R. A Irizarry, M. Åstrand, and T. P. Speed. 2003. "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias." *Bioinformatics* 19 (2): 185–93. https://doi.org/10.1093/bioinformatics/19.2.185.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Bommasani, Rishi et al. 2021. *On the Opportunities and Risks of Foundation Models*. https://arxiv.org/abs/2108.07258.

Borodovsky, Mark, and James McIninch. 1993. "GENMARK: Parallel Gene Recognition for Both DNA Strands." *Computers & Chemistry* 17 (2): 123–33. https://doi.org/10.1016/0097-8485(93)85004-v.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, et al. 2019. "BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis." *PLOS Computational Biology* 15 (4): e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent Filtering Increases Detection Power for High-Throughput Experiments." *Proceedings of the National Academy of Sciences* 107 (21): 9546–51. https://doi.org/10.1073/pnas.0914005107.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, et al. 2018. "Ten Things You Should Know about Transposable Elements." *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-018-1577-z.

Bowers, Robert M, Nikos C Kyrpides, Ramunas Stepanauskas, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31. https://doi.org/10.1038/nbt.3893.

Bowie, James U., Roland Lüthy, and David Eisenberg. 1991. "A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure." *Science* 253 (5016): 164–70. https://doi.org/10.1126/science.1853201.

Bowman, Gregory R. 2024. "AlphaFold and Protein Folding: Not Dead yet! The Frontier Is Conformational Ensembles." *Annual Review of Biomedical Data Science* 7 (1): 51–57. https://doi.org/10.1146/an

nurev-biodatasci-102423-011435.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, et al. 2008. "High-Resolution Mapping and Characterization of Open Chromatin Across the Genome." *Cell* 132 (2): 311–22. https://doi.org/10.1016/j.cell.2007.12.014.

Bradley, Philip, Kira M. S. Misura, and David Baker. 2005. "Toward High-Resolution de Novo Structure Prediction for Small Proteins." *Science* 309 (5742): 1868–71. https://doi.org/10.1126/science.1113801.

Bradnam, Keith R, Joseph N Fass, Anton Alexandrov, et al. 2013. "Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species." *Gigascience* 2 (1). https://doi.org/10.1186/2047-217x-2-10.

Brandes, Nadav, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. 2023. "Genome-Wide Prediction of Disease Variant Effects with a Deep Protein Language Model." *Nature Genetics* 55 (9): 1512–22. https://doi.org/10.1038/s41588-023-01465-0.

Branon, Tess C, Justin A Bosch, Ariana D Sanchez, et al. 2018. "Efficient Proximity Labeling in Living Cells and Organisms with TurboID." *Nature Biotechnology* 36 (9): 880–87. https://doi.org/10.1038/nbt.4201.

Braun, Efrem, Justin Gilmer, Heather B. Mayes, et al. 2018. "Best Practices for Foundations in Molecular Simulations [Article V1.0]." *Living Journal of Computational Molecular Science* 1 (1): 5957. https://doi.org/10.33011/livecoms.1.1.5957.

Braun, Pascal, Murat Tasan, Matija Dreze, et al. 2008. "An Experimentally Derived Confidence Score for Binary Protein-Protein Interactions." *Nature Methods* 6 (1): 91–97. https://doi.org/10.1038/nmeth.1281.

Bravo, Gustavo A., Alexandre Antonelli, Christine D. Bacon, et al. 2019. "Embracing Heterogeneity: Coalescing the Tree of Life and the Future of Phylogenomics." *PeerJ* 7: e6399. https://doi.org/10.7717/peerj.6399.

Bray, Mark-Anthony, Shantanu Singh, Han Han, et al. 2016. "Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes." *Nature Protocols* 11 (9): 1757–74. https://doi.org/10.1038/nprot.2016.105.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Brazma, Alvis, Pascal Hingamp, John Quackenbush, et al. 2001. "Minimum Information about a Microarray Experiment (MIAME)—Toward Standards for Microarray Data." *Nature Genetics* 29 (4): 365–71. https://doi.org/10.1038/ng1201-365.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/a:1010933404324.

Brenner, Steven E., Cyrus Chothia, and Tim J. P. Hubbard. 1998. "Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships." *Proceedings of the National Academy of Sciences* 95 (11): 6073–78. https://doi.org/10.1073/pnas.95.11.6073.

Brixi, Garyk, Matthew G. Durrant, Jerome Ku, et al. 2025. "Genome Modeling and Design Across All

Domains of Life with Evo 2." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.02.18.638918.

Broder, A. Z. 1997. "On the Resemblance and Containment of Documents." *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, SEQUEN-97, 21–29. https://doi.org/10.1109/sequen.1997.666900.

Broido, Anna D., and Aaron Clauset. 2019. "Scale-Free Networks Are Rare." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-08746-5.

Brown, Celeste J, Audra K Johnson, A Keith Dunker, and Gary W Daughdrill. 2011. "Evolution and Disorder." *Current Opinion in Structural Biology* 21 (3): 441–46. https://doi.org/10.1016/j.sbi.2011.02.005.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, et al. 2015. "Unusual Biology Across a Group Comprising More Than 15." *Nature* 523 (7559): 208–11. https://doi.org/10.1038/nature14486.

Brown, Jeremy M. 2014. "Predictive Approaches to Assessing the Fit of Evolutionary Models." *Systematic Biology* 63 (3): 289–92. https://doi.org/10.1093/sysbio/syu009.

Brůna, Tomáš, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. "BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database." *NAR Genomics and Bioinformatics* 3 (1). https://doi.org/10.1093/nargab/lqaa108.

Brünger, Axel T. 1992. "Free r Value: A Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures." *Nature* 355 (6359): 472–75. https://doi.org/10.1038/355472a0.

Brunk, Elizabeth, Swagatika Sahoo, Daniel C Zielinski, et al. 2018. "Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism." *Nature Biotechnology* 36 (3): 272–81. https://doi.org/10.1038/nbt.4072.

Bryant, D. 2003. "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks." *Molecular Biology and Evolution* 21 (2): 255–65. https://doi.org/10.1093/molbev/msh018.

Bryant, Patrick, Gabriele Pozzati, and Arne Elofsson. 2022. "Improved Prediction of Protein-Protein Interactions Using AlphaFold2." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-28865-w.

Bryant, Patrick, Gabriele Pozzati, Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. 2022. "Predicting the Structure of Large Protein Complexes Using AlphaFold and Monte Carlo Tree Search." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-33729-4.

Bryngelson, Joseph D., José Nelson Onuchic, Nicholas D. Socci, and Peter G. Wolynes. 1995. "Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis." *Proteins: Structure, Function, and Bioinformatics* 21 (3): 167–95. https://doi.org/10.1002/prot.340210302.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." *Nature Methods* 18 (4): 366–68. https://doi.org/10.1038/s41592-021-01101-x.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Budnik, Bogdan, Ezra Levy, Guillaume Harmange, and Nikolai Slavov. 2018. "SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity During Cell Differentiation." *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-018-1547-5.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, et al. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." *Nature Genetics* 47 (3): 291–95. https://doi.org/10.1038/ng.3211.

Buljan, Marija, Guilhem Chalancon, Sebastian Eustermann, et al. 2012. "Tissue-Specific Splicing of Disordered Segments That Embed Binding Motifs Rewires Protein Interaction Networks." *Molecular Cell* 46 (6): 871–83. https://doi.org/10.1016/j.molcel.2012.05.039.

Bunne, Charlotte, Yusuf Roohani, Yanay Rosen, et al. 2024. "How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities." *Cell* 187 (25): 7045–63. https://doi.org/10.1016/j.cell.2024.11.015.

Burge, Chris, and Samuel Karlin. 1997. "Prediction of Complete Gene Structures in Human Genomic DNA." *Journal of Molecular Biology* 268 (1): 78–94. https://doi.org/10.1006/jmbi.1997.0951.

Burge, S., E. Kelly, D. Lonsdale, et al. 2012. "Manual GO Annotation of Predictive Protein Signatures: The InterPro Approach to GO Curation." *Database* 2012 (0): bar068–68. https://doi.org/10.1093/database/bar068.

Burger, Lukas, and Erik van Nimwegen. 2008. "Accurate Prediction of Protein–Protein Interactions from Sequence Alignments Using a Bayesian Method." *Molecular Systems Biology* 4 (1). https://doi.org/10.1038/msb4100203.

Burke, David F., Patrick Bryant, Inigo Barrio-Hernandez, et al. 2023. "Towards a Structurally Resolved Human Protein Interaction Network." *Nature Structural & Molecular Biology* 30 (2): 216–25. https://doi.org/10.1038/s41594-022-00910-8.

Burki, Fabien, Andrew J. Roger, Matthew W. Brown, and Alastair G. B. Simpson. 2020. "The New Tree of Eukaryotes." *Trends in Ecology & Evolution* 35 (1): 43–55. https://doi.org/10.1016/j.tree.2019.08.008.

Burkowski, Forbes J. 2008. *Structural Bioinformatics: An Algorithmic Approach.* Chapman; Hall/CRC.

Burley, Stephen K., Helen M. Berman, Jose M. Duarte, et al. 2022. "Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students." *Biomolecules* 12 (10): 1425. https://doi.org/10.3390/biom12101425.

Burley, Stephen K, Helen M Berman, Charmi Bhikadiya, et al. 2018. "Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data." *Nucleic Acids Research* 47 (D1): D520–28. https://doi.org/10.1093/nar/gky949.

Burley, Stephen K., Dennis W. Piehl, Brinda Vallat, and Christine Zardecki. 2024. "RCSB Protein Data Bank: Supporting Research and Education Worldwide Through Explorations of Experimentally Determined and Computationally Predicted Atomic Level 3D Biostructures." *IUCrJ* 11 (3): 279–86. https://doi.org/10.1107/s2052252524002604.

Burset, Moisès, and Roderic Guigó. 1996. "Evaluation of Gene Structure Prediction Programs." *Genomics* 34 (3): 353–67. https://doi.org/10.1006/geno.1996.0298.

Burton, Joshua N, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman, and Jay Shendure. 2013. "Chromosome-Scale Scaffolding of de Novo Genome Assemblies Based on Chromatin Interactions." *Nature Biotechnology* 31 (12): 1119–25. https://doi.org/10.1038/nbt.2727.

Burton, Paul R., David G. Clayton, Lon R. Cardon, et al. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447 (7145): 661–78. https://doi.org/10.1038/nature05911.

Buttenschoen, Martin, Garrett M. Morris, and Charlotte M. Deane. 2024. "PoseBusters: AI-Based Docking Methods Fail to Generate Physically Valid Poses or Generalise to Novel Sequences." *Chemical Science* 15 (9): 3130–39. https://doi.org/10.1039/d3sc04185a.

Buzzao, Davide, Miguel Castresana-Aguirre, Dimitri Guala, and Erik L L Sonnhammer. 2024. "Benchmarking Enrichment Analysis Methods with the Disease Pathway Network." *Briefings in Bioinformatics* 25 (2). https://doi.org/10.1093/bib/bbae069.

Buzzao, Davide, Emma Persson, Dimitri Guala, and Erik L L Sonnhammer. 2024. "FunCoup 6: Advancing Functional Association Networks Across Species with Directed Links and Improved User Experience." *Nucleic Acids Research* 53 (D1): D658–71. https://doi.org/10.1093/nar/gkae1021.

Caicedo, Juan C, Sam Cooper, Florian Heigwer, et al. 2017. "Data-Analysis Strategies for Image-Based Cell Profiling." *Nature Methods* 14 (9): 849–63. https://doi.org/10.1038/nmeth.4397.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Candido, Salvatore, Thomas Hayes, Alexander Derry, et al. 2026. "Language Modeling Materializes a World Model of Protein Biology." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.64898/2026.06.03.729735.

Cann, Rebecca L., Mark Stoneking, and Allan C. Wilson. 1987. "Mitochondrial DNA and Human Evolution." *Nature* 325 (6099): 31–36. https://doi.org/10.1038/325031a0.

Cantalapiedra, Carlos P, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. "eggNOG-Mapper V2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38 (12): 5825–29. https://doi.org/10.1093/molbev/msab293.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Caprioli, Richard M., Terry B. Farmer, and Jocelyn Gile. 1997. "Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS." *Analytical Chemistry* 69 (23): 4751–60. https://doi.org/10.1021/ac970888i.

Carlson, Christopher S., Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. 2004. "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium." *The American Journal of Human Genetics* 74 (1): 106–20. https://doi.org/10.1086/381000.

Caron, Mathilde, Hugo Touvron, Ishan Misra, et al. 2021. "Emerging Properties in Self-Supervised Vision Transformers." *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–40. https://doi.org/10.1109/iccv48922.2021.00951.

Carpenter, Anne E, Thouis R Jones, Michael R Lamprecht, et al. 2006. "CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes." *Genome Biology* 7 (10). https://doi.org/10.1186/gb-2006-7-10-r100.

Carrillo, Humberto, and David Lipman. 1988. "The Multiple Sequence Alignment Problem in Biology." *SIAM Journal on Applied Mathematics* 48 (5): 1073–82. https://doi.org/10.1137/0148063.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, et al. 2020. "The CARE Principles for Indigenous Data Governance." *Data Science Journal* 19. https://doi.org/10.5334/dsj-2020-043.

Carter, Scott L, Kristian Cibulskis, Elena Helman, et al. 2012. "Absolute Quantification of Somatic DNA Alterations in Human Cancer." *Nature Biotechnology* 30 (5): 413–21. https://doi.org/10.1038/nbt.2203.

Carvalho, Claudia M. B., and James R. Lupski. 2016. "Mechanisms Underlying Structural Variant Formation in Genomic Disorders." *Nature Reviews Genetics* 17 (4): 224–38. https://doi.org/10.1038/nrg.2015.25.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, et al. 2012. "Proto-Genes and de Novo Gene Birth." *Nature* 487 (7407): 370–74. https://doi.org/10.1038/nature11184.

Casillas, Sònia, and Antonio Barbadilla. 2017. "Molecular Population Genetics." *Genetics* 205 (3): 1003–35. https://doi.org/10.1534/genetics.116.196493.

Castresana-Aguirre, Miguel, Emma Persson, and Erik L L Sonnhammer. 2021. "PathBIX—a Web Server for Network-Based Pathway Annotation with Adaptive Null Models." *Bioinformatics Advances* 1 (1). https://doi.org/10.1093/bioadv/vbab010.

Castresana-Aguirre, Miguel, and Erik L. L. Sonnhammer. 2020. "Pathway-Specific Model Estimation for Improved Pathway Annotation by Network Crosstalk." *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-70239-z.

Chamier, Lucas von, Romain F. Laine, and Ricardo Henriques. 2019. "Artificial Intelligence for Microscopy: What You Should Know." *Biochemical Society Transactions* 47 (4): 1029–40. https://doi.org/10.1042/bst20180391.

Chamier, Lucas von, Romain F. Laine, Johanna Jukkala, et al. 2021. "Democratising Deep Learning for Microscopy with ZeroCostDL4Mic." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-22518-0.

Chan, Patricia P, Brian Y Lin, Allysia J Mak, and Todd M Lowe. 2021. "tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes." *Nucleic Acids Research* 49 (16): 9077–96. https://doi.org/10.1093/nar/gkab688.

Chandra, Abel, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. 2023. "Transformer-Based Deep Learning for Predicting Protein Properties in the Life Sciences." *eLife* 12. https://doi.org/10.7554/elife.82819.

Chandrasekaran, Srinivas Niranj, Eric Alix, John Arevalo, et al. 2025. "Morphological Map of Under- and Overexpression of Genes in Human Cells." *Nature Methods* 22 (8): 1742–52. https://doi.org/10.1038/s41592-025-02753-9.

Chandrasekaran, Srinivas Niranj, Beth A. Cimini, Amy Goodale, et al. 2024. "Three Million Images and Morphological Profiles of Cells Treated with Matched Chemical and Genetic Perturbations." *Nature Methods* 21 (6): 1114–21. https://doi.org/10.1038/s41592-024-02241-6.

Chang, Antje, Lisa Jeske, Sandra Ulbrich, et al. 2020. "BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates." *Nucleic Acids Research* 49 (D1): D498–508. https://doi.org/10.1093/nar/gkaa1025.

Chapman, Matthew R., Lloyd S. Robinson, Jerome S. Pinkner, et al. 2002. "Role of Escherichia Coli Curli Operons in Directing Amyloid Fiber Formation." *Science* 295 (5556): 851–55. https://doi.org/10.1126/science.1067484.

Chari, Tara, and Lior Pachter. 2023. "The Specious Art of Single-Cell Genomics." *PLOS Computational Biology* 19 (8): e1011288. https://doi.org/10.1371/journal.pcbi.1011288.

Charlesworth, Brian. 2009. "Effective Population Size and Patterns of Molecular Evolution and Variation." *Nature Reviews Genetics* 10 (3): 195–205. https://doi.org/10.1038/nrg2526.

Cheetham, Seth W., Geoffrey J. Faulkner, and Marcel E. Dinger. 2019. "Overcoming Challenges and Dogmas to Understand the Functions of Pseudogenes." *Nature Reviews Genetics* 21 (3): 191–201. https://doi.org/10.1038/s41576-019-0196-1.

Chen, Jian-Min, David N. Cooper, Nadia Chuzhanova, Claude Férec, and George P. Patrinos. 2007. "Gene Conversion: Mechanisms, Evolution and Human Disease." *Nature Reviews Genetics* 8 (10): 762–75. https://doi.org/10.1038/nrg2193.

Chen, Jin, Andreas-David Brunner, J. Zachery Cogan, et al. 2020. "Pervasive Functional Translation of Noncanonical Human Open Reading Frames." *Science* 367 (6482): 1140–46. https://doi.org/10.1126/science.aay0262.

Chen, Kok Hao, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. 2015. "Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells." *Science* 348 (6233). https://doi.org/10.1126/science.aaa6090.

Chen, Richard J., Tong Ding, Ming Y. Lu, et al. 2024. "Towards a General-Purpose Foundation Model for Computational Pathology." *Nature Medicine* 30 (3): 850–62. https://doi.org/10.1038/s41591-024-02857-3.

Chen, Siwei, Laurent C. Francioli, Julia K. Goodrich, et al. 2023. "A Genomic Mutational Constraint Map

Using Variation in 76,156 Human Genomes." *Nature* 625 (7993): 92–100. https://doi.org/10.1038/s41586-023-06045-0.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–94. https://doi.org/10.1145/2939672.2939785.

Chen, Xinshi, Yuxuan Zhang, Chan Lu, et al. 2025. "Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.01.08.631967.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm." *Nature Methods* 18 (2): 170–75. https://doi.org/10.1038/s41592-020-01056-5.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, et al. 2014. "ECOD: An Evolutionary Classification of Protein Domains." *PLoS Computational Biology* 10 (12): e1003926. https://doi.org/10.1371/journal.pcbi.1003926.

Cheng, Jun, Guido Novati, Joshua Pan, et al. 2023. "Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense." *Science* 381 (6664). https://doi.org/10.1126/science.adg7492.

Chicco, Davide. 2017. "Ten Quick Tips for Machine Learning in Computational Biology." *BioData Mining* 10 (1). https://doi.org/10.1186/s13040-017-0155-3.

Chicco, Davide, and Giuseppe Agapito. 2022. "Nine Quick Tips for Pathway Enrichment Analysis." *PLOS Computational Biology* 18 (8): e1010348. https://doi.org/10.1371/journal.pcbi.1010348.

Chicco, Davide, and Giuseppe Jurman. 2020. "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation." *BMC Genomics* 21 (1). https://doi.org/10.1186/s12864-019-6413-7.

Chin, Chen-Shan, David H Alexander, Patrick Marks, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nature Methods* 10 (6): 563–69. https://doi.org/10.1038/nmeth.2474.

Ching, Travers, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, et al. 2018. "Opportunities and Obstacles for Deep Learning in Biology and Medicine." *Journal of The Royal Society Interface* 15 (141). https://doi.org/10.1098/rsif.2017.0387.

Chiti, Fabrizio, and Christopher M. Dobson. 2017. "Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress over the Last Decade." *Annual Review of Biochemistry* 86 (1): 27–68. https://doi.org/10.1146/annurev-biochem-061516-045115.

Cho, Nathan H., Keith C. Cheveralls, Andreas-David Brunner, et al. 2022. "OpenCell: Endogenous Tagging for the Cartography of Human Cellular Organization." *Science* 375 (6585). https://doi.org/10.1126/science.abi6983.

Cho, Yehlin, Griffin Rangel, Gaurav Bhardwaj, and Sergey Ovchinnikov. 2025. "Protein Hunter: Exploiting Structure Hallucination Within Diffusion for Protein Design." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.10.10.681530.

Choi, Hyungwon, Brett Larsen, Zhen-Yuan Lin, et al. 2010. "SAINT: Probabilistic Scoring of Affinity Purification–Mass Spectrometry Data." *Nature Methods* 8 (1): 70–73. https://doi.org/10.1038/nmeth.1541.

Choi, Meena, Ching-Yun Chang, Timothy Clough, et al. 2014. "MSstats: An r Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments." *Bioinformatics* 30 (17): 2524–26. https://doi.org/10.1093/bioinformatics/btu305.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O'Reilly. 2020. "Tutorial: A Guide to Performing Polygenic Risk Score Analyses." *Nature Protocols* 15 (9): 2759–72. https://doi.org/10.1038/s41596-020-0353-1.

Chothani, Sonia P., Eleonora Adami, Anissa A. Widjaja, et al. 2022. "A High-Resolution Map of Human RNA Translation." *Molecular Cell* 82 (15): 2885–2899.e8. https://doi.org/10.1016/j.molcel.2022.06.023.

Chothia, C., and A. M. Lesk. 1986. "The Relation Between the Divergence of Sequence and Structure in Proteins." *The EMBO Journal* 5 (4): 823–26. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

Chothia, C, M Levitt, and D Richardson. 1977. "Structure of Proteins: Packing of Alpha-Helices and Pleated Sheets." *Proceedings of the National Academy of Sciences* 74 (10): 4130–34. https://doi.org/10.1073/pnas.74.10.4130.

Chothia, Cyrus. 1992. "One Thousand Families for the Molecular Biologist." *Nature* 357 (6379): 543–44. https://doi.org/10.1038/357543a0.

Chothia, Cyrus, and Joël Janin. 1975. "Principles of Protein–Protein Recognition." *Nature* 256 (5520): 705–8. https://doi.org/10.1038/256705a0.

Chou, Peter Y., and Gerald D. Fasman. 1974a. "Conformational Parameters for Amino Acids in Helical, β-Sheet, and Random Coil Regions Calculated from Proteins." *Biochemistry* 13 (2): 211–22. https://doi.org/10.1021/bi00699a001.

Chou, Peter Y., and Gerald D. Fasman. 1974b. "Prediction of Protein Conformation." *Biochemistry* 13 (2): 222–45. https://doi.org/10.1021/bi00699a002.

Chow, Louise T., Richard E. Gelinas, Thomas R. Broker, and Richard J. Roberts. 1977. "An Amazing Sequence Arrangement at the 5′ Ends of Adenovirus 2 Messenger RNA." *Cell* 12 (1): 1–8. https://doi.org/10.1016/0092-8674(77)90180-5.

Christiansen, Eric M., Samuel J. Yang, D. Michael Ando, et al. 2018. "In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images." *Cell* 173 (3): 792–803.e19. https://doi.org/10.1016/j.cell.2018.03.040.

Christmas, Matthew J., Irene M. Kaplow, Diane P. Genereux, et al. 2023. "Evolutionary Constraint and Innovation Across Hundreds of Placental Mammals." *Science* 380 (6643). https://doi.org/10.1126/science.abn3943.

Christoforou, Andy, Claire M. Mulvey, Lisa M. Breckels, et al. 2016. "A Draft Map of the Mouse Pluripotent Stem Cell Spatial Proteome." *Nature Communications* 7 (1). https://doi.org/10.1038/ncomms9992.

Churchill, Gary A. 1989. "Stochastic Models for Heterogeneous DNA Sequences." *Bulletin of Mathematical Biology* 51 (1): 79–94. https://doi.org/10.1007/bf02458837.

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, et al. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology* 31 (3): 213–19. https://doi.org/10.1038/nbt.2514.

Cimini, Beth A., Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, et al. 2023. "Optimizing the Cell Painting Assay for Image-Based Profiling." *Nature Protocols* 18 (7): 1981–2013. https://doi.org/10.1038/s41596-023-00840-9.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, et al. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w 1118 ; Iso-2; Iso-3." *Fly* 6 (2): 80–92. https://doi.org/10.4161/fly.19695.

Clamp, Michele, Ben Fry, Mike Kamal, et al. 2007. "Distinguishing Protein-Coding and Noncoding Genes in the Human Genome." *Proceedings of the National Academy of Sciences* 104 (49): 19428–33. https://doi.org/10.1073/pnas.0709013104.

Clark, Douglas, Choi (OpenStax). 2018a. *Chapter 14: DNA Structure and Function (Biology 2e).* OpenStax Biology 2e. https://openstax.org/books/biology-2e/pages/14-introduction.

Clark, Douglas, Choi (OpenStax). 2018b. *Chapter 15: Genes and Proteins (Biology 2e).* OpenStax Biology 2e. https://openstax.org/books/biology-2e/pages/15-introduction.

Clark, Mary Ann, Matthew Douglas, and Jung Choi. 2018. *Biology 2e.* OpenStax. https://openstax.org/details/books/biology-2e.

Clark, Wyatt T., and Predrag Radivojac. 2013. "Information-Theoretic Evaluation of Predicted Ontological Annotations." *Bioinformatics* 29 (13): i53–61. https://doi.org/10.1093/bioinformatics/btt228.

Clough, Emily, Tanya Barrett, Stephen E Wilhite, et al. 2023. "NCBI GEO: Archive for Gene Expression and Epigenomics Data Sets: 23-Year Update." *Nucleic Acids Research* 52 (D1): D138–44. https://doi.org/10.1093/nar/gkad965.

Coleman, Gareth A., Adrián A. Davín, Tara A. Mahendrarajah, et al. 2021. "A Rooted Phylogeny Resolves Early Bacterial Evolution." *Science* 372 (6542). https://doi.org/10.1126/science.abe0511.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, et al. 2020. "A Structural Variation Reference for Medical and Population Genetics." *Nature* 581 (7809): 444–51. https://doi.org/10.1038/s41586-020-2287-8.

Compeau, Phillip E C, Pavel A Pevzner, and Glenn Tesler. 2011. "How to Apply de Bruijn Graphs to Genome Assembly." *Nature Biotechnology* 29 (11): 987–91. https://doi.org/10.1038/nbt.2023.

Compeau, Phillip, and Pavel Pevzner. 2018. *Bioinformatics Algorithms: An Active Learning Approach.* 3rd ed. Active Learning Publishers. https://www.bioinformaticsalgorithms.org/.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0881-8.

Cong, Qian, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. 2019. "Protein Interaction Networks Revealed by Proteome Coevolution." *Science* 365 (6449): 185–89. https://doi.org/10.1126/science.aaw6718.

Cooper, Gregory M., Eric A. Stone, George Asimenos, Eric D. Green, Serafim Batzoglou, and Arend Sidow. 2005. "Distribution and Intensity of Constraint in Mammalian Genomic Sequence." *Genome Research* 15 (7): 901–13. https://doi.org/10.1101/gr.3577405.

Cornish-Bowden, Athel. 2015. "One Hundred Years of Michaelis–Menten Kinetics." *Perspectives in Science* 4: 3–9. https://doi.org/10.1016/j.pisc.2014.12.002.

Corpas, Manuel, Nadezda V. Kovalevskaya, Amanda McMurray, and Fiona G. G. Nielsen. 2018. "A FAIR Guide for Data Providers to Maximise Sharing of Human Genomic Data." *PLOS Computational Biology* 14 (3): e1005873. https://doi.org/10.1371/journal.pcbi.1005873.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97. https://doi.org/10.1007/bf00994018.

Costanzo, Michael, Benjamin VanderSluis, Elizabeth N. Koch, et al. 2016. "A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function." *Science* 353 (6306). https://doi.org/10.1126/science.aaf1420.

Couso, Juan-Pablo, and Pedro Patraquim. 2017. "Classification and Function of Small Open Reading Frames." *Nature Reviews Molecular Cell Biology* 18 (9): 575–89. https://doi.org/10.1038/nrm.2017.58.

Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. "Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ." *Molecular & Cellular Proteomics* 13 (9): 2513–26. https://doi.org/10.1074/mcp.m113.031591.

Cox, Jürgen, and Matthias Mann. 2008. "MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification." *Nature Biotechnology* 26 (12): 1367–72. https://doi.org/10.1038/nbt.1511.

Cox, Jürgen, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. 2011. "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment." *Journal of Proteome Research* 10 (4): 1794–805. https://doi.org/10.1021/pr101065j.

Cozzetto, Domenico, and David T. Jones. 2016. "Computational Methods for Annotation Transfers from Sequence." In *The Gene Ontology Handbook*. Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_5.

Craig, Robertson, and Ronald C. Beavis. 2004. "TANDEM: Matching Proteins with Tandem Mass Spectra." *Bioinformatics* 20 (9): 1466–67. https://doi.org/10.1093/bioinformatics/bth092.

Crowell, Helena L., Charlotte Soneson, Pierre-Luc Germain, et al. 2020. "Muscat Detects Subpopulation-Specific State Transitions from Multi-Sample Multi-Condition Single-Cell Transcriptomics Data." *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-020-19894-4.

Cui, Haotian, Chloe Wang, Hassaan Maan, et al. 2024. "scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI." *Nature Methods* 21 (8): 1470–80. https://doi.org/10.1

038/s41592-024-02201-0.

D'haeseleer, Patrik. 2005. "How Does Gene Expression Clustering Work?" *Nature Biotechnology* 23 (12): 1499–501. https://doi.org/10.1038/nbt1205-1499.

Dahiyat, Bassil I., and Stephen L. Mayo. 1997. "De Novo Protein Design: Fully Automated Sequence Selection." *Science* 278 (5335): 82–87. https://doi.org/10.1126/science.278.5335.82.

Dalal, Seema, Suganthi Balasubramanian, and Lynne Regan. 1997. "Protein Alchemy: Changing β-Sheet into α-Helix." *Nature Structural Biology* 4 (7): 548–52. https://doi.org/10.1038/nsb0797-548.

Dalla-Torre, Hugo, Liam Gonzalez, Javier Mendoza-Revilla, et al. 2024. "Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics." *Nature Methods* 22 (2): 287–97. https://doi.org/10.1038/s41592-024-02523-z.

Dam, Sipko van, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. 2017. "Gene Co-Expression Analysis for Functional Classification and Gene–Disease Predictions." *Briefings in Bioinformatics*, bbw139. https://doi.org/10.1093/bib/bbw139.

Dančík, Vlado, Theresa A. Addona, Karl R. Clauser, James E. Vath, and Pavel A. Pevzner. 1999. "De Novo Peptide Sequencing via Tandem Mass Spectrometry." *Journal of Computational Biology* 6 (3-4): 327–42. https://doi.org/10.1089/106652799318300.

Dandekar, T. 1998. "Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact." *Trends in Biochemical Sciences* 23 (9): 324–28. https://doi.org/10.1016/s0968-0004(98)01274-2.

Danecek, Petr, Adam Auton, Goncalo Abecasis, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Danecek, Petr, James K Bonfield, Jennifer Liddle, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Danielewski, Mikołaj, Marlena Szalata, Jan Krzysztof Nowak, Jarosław Walkowiak, Ryszard Słomski, and Karolina Wielgus. 2025. "History of Biological Databases, Their Importance, and Existence in Modern Scientific and Policy Context." *Genes* 16 (1): 100. https://doi.org/10.3390/genes16010100.

Dann, Emma, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. 2021. "Differential Abundance Testing on Single-Cell Data Using k-Nearest Neighbor Graphs." *Nature Biotechnology* 40 (2): 245–53. https://doi.org/10.1038/s41587-021-01033-z.

Darling, Aaron C. E., Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. "Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements." *Genome Research* 14 (7): 1394–403. https://doi.org/10.1101/gr.2289704.

Das, Sayoni, Natalie L Dawson, and Christine A Orengo. 2015. "Diversity in Protein Domain Superfamilies." *Current Opinion in Genetics & Development* 35: 40–49. https://doi.org/10.1016/j.gde.2015.09.005.

Dauparas, J., I. Anishchenko, N. Bennett, et al. 2022. "Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN." *Science* 378 (6615): 49–56. https://doi.org/10.1126/science.add2187.

Davey Smith, George, and Shah Ebrahim. 2003. "'Mendelian Randomization': Can Genetic Epidemiology

Contribute to Understanding Environmental Determinants of Disease?*." *International Journal of Epidemiology* 32 (1): 1–22. https://doi.org/10.1093/ije/dyg070.

David, Alessia, Suhail Islam, Evgeny Tankhilevich, and Michael J. E. Sternberg. 2022. "The AlphaFold Database of Protein Structures: A Biologist's Guide." *Journal of Molecular Biology* 434 (2): 167336. https://doi.org/10.1016/j.jmb.2021.167336.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be Under Selective Constraint Using GERP++." *PLoS Computational Biology* 6 (12): e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. "A Model of Evolutionary Change in Proteins." In *Atlas of Protein Sequence and Structure*, vol. 5. National Biomedical Research Foundation.

De Bie, Tijl, Nello Cristianini, Jeffery P. Demuth, and Matthew W. Hahn. 2006. "CAFE: A Computational Tool for the Study of Gene Family Evolution." *Bioinformatics* 22 (10): 1269–71. https://doi.org/10.1093/bioinformatics/btl097.

De Las Rivas, Javier, and Celia Fontanillo. 2010. "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks." *PLoS Computational Biology* 6 (6): e1000807. https://doi.org/10.1371/journal.pcbi.1000807.

De Paolis Kaluza, M. Clara, Rashika Ramola, Parnal Joshi, et al. 2026. "Advances in Protein Function Prediction from the Fifth CAFA Challenge." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.64898/2026.04.27.716980.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. "Three Decades of Nanopore Sequencing." *Nature Biotechnology* 34 (5): 518–24. https://doi.org/10.1038/nbt.3423.

Degnan, James H., and Noah A. Rosenberg. 2009. "Gene Tree Discordance, Phylogenetic Inference and the Multispecies Coalescent." *Trends in Ecology & Evolution* 24 (6): 332–40. https://doi.org/10.1016/j.tree.2009.01.009.

Degnan, James H, and Noah A Rosenberg. 2006. "Discordance of Species Trees with Their Most Likely Gene Trees." *PLoS Genetics* 2 (5): e68. https://doi.org/10.1371/journal.pgen.0020068.

Dehal, Paramvir, and Jeffrey L Boore. 2005. "Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate." *PLoS Biology* 3 (10): e314. https://doi.org/10.1371/journal.pbio.0030314.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." *Science* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13225-y.

Delcher, A. 1999. "Improved Microbial Gene Identification with GLIMMER." *Nucleic Acids Research* 27 (23): 4636–41. https://doi.org/10.1093/nar/27.23.4636.

Delsuc, Frédéric, Henner Brinkmann, and Hervé Philippe. 2005. "Phylogenomics and the Reconstruction

of the Tree of Life." *Nature Reviews Genetics* 6 (5): 361–75. https://doi.org/10.1038/nrg1603.

Demichev, Vadim, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. 2019. "DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput." *Nature Methods* 17 (1): 41–44. https://doi.org/10.1038/s41592-019-0638-x.

Dempster, Joshua M., Isabella Boyle, Francisca Vazquez, et al. 2021. "Chronos: A Cell Population Dynamics Model of CRISPR Experiments That Improves Inference of Gene Fitness Effects." *Genome Biology* 22 (1). https://doi.org/10.1186/s13059-021-02540-7.

Demuth, Jeffery P., Tijl De Bie, Jason E. Stajich, Nello Cristianini, and Matthew W. Hahn. 2006. "The Evolution of Mammalian Gene Families." *PLoS ONE* 1 (1): e85. https://doi.org/10.1371/journal.pone.0000085.

DePristo, Mark A, Eric Banks, Ryan Poplin, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

Des Marais, David J., Joseph A. Nuth, Louis J. Allamandola, et al. 2008. "The NASA Astrobiology Roadmap." *Astrobiology* 8 (4): 715–30. https://doi.org/10.1089/ast.2008.0819.

Desiere, F. 2006. "The PeptideAtlas Project." *Nucleic Acids Research* 34 (90001): D655–58. https://doi.org/10.1093/nar/gkj040.

Desmet, Johan, Marc De Maeyer, Bart Hazes, and Ignace Lasters. 1992. "The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning." *Nature* 356 (6369): 539–42. https://doi.org/10.1038/356539a0.

Desper, Richard, and Olivier Gascuel. 2002. "Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle." *Journal of Computational Biology* 9 (5): 687–705. https://doi.org/10.1089/106652702761034136.

Deutsch, Eric W, Nuno Bandeira, Yasset Perez-Riverol, et al. 2022. "The ProteomeXchange Consortium at 10 Years: 2023 Update." *Nucleic Acids Research* 51 (D1): D1539–48. https://doi.org/10.1093/nar/gkac1040.

Deutsch, Eric W., Juan Antonio Vizcaíno, Andrew R. Jones, et al. 2023. "Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work." *Journal of Proteome Research* 22 (2): 287–301. https://doi.org/10.1021/acs.jproteome.2c00637.

Devlin, B., and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x.

Devos, Damien, and Alfonso Valencia. 2000. "Practical Limits of Function Prediction." *Proteins: Structure, Function, and Genetics* 41 (1): 98–107. https://doi.org/10.1002/1097-0134(20001001)41:1%3C98::aid-prot120%3E3.0.co;2-s.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Dickinson, Mary E. et al. 2016. "High-Throughput Discovery of Novel Developmental Phenotypes." *Nature* 537 (7621): 508–14. https://doi.org/10.1038/nature19356.

Dill, Ken A., and Hue Sun Chan. 1997. "From Levinthal to Pathways to Funnels." *Nature Structural & Molecular Biology* 4 (1): 10–19. https://doi.org/10.1038/nsb0197-10.

Dill, Ken A., and Justin L. MacCallum. 2012. "The Protein-Folding Problem, 50 Years On." *Science* 338 (6110): 1042–46. https://doi.org/10.1126/science.1219021.

Dixit, Atray, Oren Parnas, Biyu Li, et al. 2016. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." *Cell* 167 (7): 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, et al. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485 (7398): 376–80. https://doi.org/10.1038/nature11082.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, et al. 2012. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dodd, Matthew S., Dominic Papineau, Tor Grenne, et al. 2017. "Evidence for Early Life in Earth's Oldest Hydrothermal Vent Precipitates." *Nature* 543 (7643): 60–64. https://doi.org/10.1038/nature21377.

Dominguez, Cyril, Rolf Boelens, and Alexandre M. J. J. Bonvin. 2003. "HADDOCK: A Protein–protein Docking Approach Based on Biochemical or Biophysical Information." *Journal of the American Chemical Society* 125 (7): 1731–37. https://doi.org/10.1021/ja026939x.

Dominguez Del Angel, Victoria, Erik Hjerde, Lieven Sterck, et al. 2018. "Ten Steps to Get Started in Genome Assembly and Annotation." *F1000Research* 7: 148. https://doi.org/10.12688/f1000research.13598.1.

Doolittle, Russell F. 1999. "Do You Dig My Groove?" *Nature Genetics* 23 (1): 6–8. https://doi.org/10.1038/12597.

Doolittle, Russell F. 1981. "Similar Amino Acid Sequences: Chance or Common Ancestry?" *Science* 214 (4517): 149–59. https://doi.org/10.1126/science.7280687.

Doolittle, Russell F. 2010. "The Roots of Bioinformatics in Protein Evolution." *PLoS Computational Biology* 6 (7): e1000875. https://doi.org/10.1371/journal.pcbi.1000875.

Doolittle, Russell F., Michael W. Hunkapiller, Leroy E. Hood, et al. 1983. "Simian Sarcoma Virus Onc Gene, v- Sis , Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor." *Science* 221 (4607): 275–77. https://doi.org/10.1126/science.6304883.

Doolittle, W. Ford. 1999. "Phylogenetic Classification and the Universal Tree." *Science* 284 (5423): 2124–28. https://doi.org/10.1126/science.284.5423.2124.

Doolittle, W. Ford. 2013. "Is Junk DNA Bunk? A Critique of ENCODE." *Proceedings of the National Academy of Sciences* 110 (14): 5294–300. https://doi.org/10.1073/pnas.1221376110.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2021. *An Image Is Worth 16x16 Words:*

Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929.

Drew, Kevin, Chanjae Lee, Ryan L Huizar, et al. 2017. "Integration of over 9,000 Mass Spectrometry Experiments Builds a Global Map of Human Protein Complexes." *Molecular Systems Biology* 13 (6). https://doi.org/10.15252/msb.20167490.

Drew, Kevin, John B Wallingford, and Edward M Marcotte. 2021. "Hu.MAP 2.0: Integration of over 15,000 Proteomic Experiments Builds a Global Compendium of Human Multiprotein Assemblies." *Molecular Systems Biology* 17 (5). https://doi.org/10.15252/msb.202010016.

Drummond, Alexei J, Simon Y. W Ho, Matthew J Phillips, and Andrew Rambaut. 2006. "Relaxed Phylogenetics and Dating with Confidence." *PLoS Biology* 4 (5): e88. https://doi.org/10.1371/journal.pbio.0040088.

Drummond, D. Allan, Jesse D. Bloom, Christoph Adami, Claus O. Wilke, and Frances H. Arnold. 2005. "Why Highly Expressed Proteins Evolve Slowly." *Proceedings of the National Academy of Sciences* 102 (40): 14338–43. https://doi.org/10.1073/pnas.0504070102.

Dunbrack, Roland L., and Martin Karplus. 1993. "Backbone-Dependent Rotamer Library for Proteins Application to Side-Chain Prediction." *Journal of Molecular Biology* 230 (2): 543–74. https://doi.org/10.1006/jmbi.1993.1170.

Dupree, Emmalyn J., Madhuri Jayathirtha, Hannah Yorkey, Marius Mihasan, Brindusa Alina Petre, and Costel C. Darie. 2020. "A Critical Review of Bottom-up Proteomics: The Good, the Bad, and the Future of This Field." *Proteomes* 8 (3): 14. https://doi.org/10.3390/proteomes8030014.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. "Testing for Ancient Admixture Between Closely Related Populations." *Molecular Biology and Evolution* 28 (8): 2239–52. https://doi.org/10.1093/molbev/msr048.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. https://doi.org/10.1017/cbo9780511790492.

Dyer, Sarah C, Olanrewaju Austine-Orimoloye, Andrey G Azov, et al. 2024. "Ensembl 2025." *Nucleic Acids Research* 53 (D1): D948–57. https://doi.org/10.1093/nar/gkae1071.

Earl, Dent, Ngan Nguyen, Glenn Hickey, et al. 2014. "Alignathon: A Competitive Assessment of Whole-Genome Alignment Methods." *Genome Research* 24 (12): 2077–89. https://doi.org/10.1101/gr.174920.114.

Echave, Julian, Stephanie J. Spielman, and Claus O. Wilke. 2016. "Causes of Evolutionary Rate Variation Among Protein Sites." *Nature Reviews Genetics* 17 (2): 109–21. https://doi.org/10.1038/nrg.2015.18.

Eddy, S R. 1998. "Profile Hidden Markov Models." *Bioinformatics* 14 (9): 755–63. https://doi.org/10.1093/bioinformatics/14.9.755.

Eddy, Sean R. 2004a. "What Is a Hidden Markov Model?" *Nature Biotechnology* 22 (10): 1315–16. https://doi.org/10.1038/nbt1004-1315.

Eddy, Sean R. 2004b. "Where Did the BLOSUM62 Alignment Score Matrix Come From?" *Nature*

Biotechnology 22 (8): 1035–36. https://doi.org/10.1038/nbt0804-1035.

Eddy, Sean R. 2008. "A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation." *PLoS Computational Biology* 4 (5): e1000069. https://doi.org/10.1371/journal.pcbi.1000069.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Eddy, Sean R. 2012. "The c-Value Paradox, Junk DNA and ENCODE." *Current Biology* 22 (21): R898–99. https://doi.org/10.1016/j.cub.2012.10.002.

Eddy, Sean R., and Richard Durbin. 1994. "RNA Sequence Analysis Using Covariance Models." *Nucleic Acids Research* 22 (11): 2079–88. https://doi.org/10.1093/nar/22.11.2079.

Edfors, Fredrik, Andreas Hober, Klas Linderbäck, et al. 2018. "Enhanced Validation of Antibodies for Research Applications." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-06642-y.

Edgar, R. C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Edgar, Robert C. 2022. "Muscle5: High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-34630-w.

Edwards, J. S., and B. O. Palsson. 2000. "The Escherichia Coli MG1655 in Silico Metabolic Genotype: Its Definition, Characteristics, and Capabilities." *Proceedings of the National Academy of Sciences* 97 (10): 5528–33. https://doi.org/10.1073/pnas.97.10.5528.

Edwards, Scott V. 2009. "IS a NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING?" *Evolution* 63 (1): 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x.

Efron, Bradley, and Robert Tibshirani. 2002. "Empirical Bayes Methods and False Discovery Rates for Microarrays." *Genetic Epidemiology* 23 (1): 70–86. https://doi.org/10.1002/gepi.1124.

Eid, John, Adrian Fehr, Jeremy Gray, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Eisen, Jonathan A. 1998. "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis." *Genome Research* 8 (3): 163–67. https://doi.org/10.1101/gr.8.3.163.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein. 1998. "Cluster Analysis and Display of Genome-Wide Expression Patterns." *Proceedings of the National Academy of Sciences* 95 (25): 14863–68. https://doi.org/10.1073/pnas.95.25.14863.

Ejigu, Girum Fitihamlak, and Jaehee Jung. 2020. "Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing." *Biology* 9 (9): 295. https://doi.org/10.3390/biology9090295.

Ekblom, Robert, and Jochen B. W. Wolf. 2014. "A Field Guide to Whole-Genome Sequencing, Assembly and Annotation." *Evolutionary Applications* 7 (9): 1026–42. https://doi.org/10.1111/eva.12178.

Ekman, Diana, Sara Light, Åsa K Björklund, and Arne Elofsson. 2006. "What Properties Characterize the Hub Proteins of the Protein-Protein Interaction Network of Saccharomyces Cerevisiae?" *Genome Biology* 7 (6). https://doi.org/10.1186/gb-2006-7-6-r45.

Eldjarn, Grimur Hjorleifsson, Egil Ferkingstad, Sigrun H. Lund, et al. 2023. "Large-Scale Plasma Proteomics Comparisons Through Genetics and Disease Associations." *Nature* 622 (7982): 348–58. https://doi.org/10.1038/s41586-023-06563-x.

Elias, Joshua E, and Steven P Gygi. 2007. "Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry." *Nature Methods* 4 (3): 207–14. https://doi.org/10.1038/nmeth1019.

Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, et al. 2022. "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (10): 7112–27. https://doi.org/10.1109/tpami.2021.3095381.

Eloff, Kevin, Konstantinos Kalogeropoulos, Amandla Mabona, et al. 2025. "InstaNovo Enables Diffusion-Powered de Novo Peptide Sequencing in Large-Scale Proteomics Experiments." *Nature Machine Intelligence* 7 (4): 565–79. https://doi.org/10.1038/s42256-025-01019-5.

Elofsson, Arne. 2023. "Progress at Protein Structure Prediction, as Seen in CASP15." *Current Opinion in Structural Biology* 80: 102594. https://doi.org/10.1016/j.sbi.2023.102594.

Elofsson, Arne, and Gunnar von Heijne. 2007. "Membrane Protein Structure: Prediction Versus Reality." *Annual Review of Biochemistry* 76 (1): 125–40. https://doi.org/10.1146/annurev.biochem.76.052705.163539.

Elowitz, Michael B., and Stanislas Leibler. 2000. "A Synthetic Oscillatory Network of Transcriptional Regulators." *Nature* 403 (6767): 335–38. https://doi.org/10.1038/35002125.

Elowitz, Michael B., Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. 2002. "Stochastic Gene Expression in a Single Cell." *Science* 297 (5584): 1183–86. https://doi.org/10.1126/science.1070919.

Emanuelsson, Olof, Henrik Nielsen, Søren Brunak, and Gunnar von Heijne. 2000. "Predicting Subcellular Localization of Proteins Based on Their n-Terminal Amino Acid Sequence." *Journal of Molecular Biology* 300 (4): 1005–16. https://doi.org/10.1006/jmbi.2000.3903.

Eme, Laura, Daniel Tamarit, Eva F. Caceres, et al. 2023. "Inference and Reconstruction of the Heimdallarchaeial Ancestry of Eukaryotes." *Nature* 618 (7967): 992–99. https://doi.org/10.1038/s41586-023-06186-2.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1832-y.

Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. 1994. "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database." *Journal of the American Society for Mass Spectrometry* 5 (11): 976–89. https://doi.org/10.1016/1044-0305(94)80016-2.

Engelhardt, Barbara E, Michael I Jordan, Kathryn E Muratore, and Steven E Brenner. 2005. "Protein Molecular Function Prediction by Bayesian Phylogenomics." *PLoS Computational Biology* 1 (5): e45. https://doi.org/10.1371/journal.pcbi.0010045.

Enright, A. J. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84. https://doi.org/10.1093/nar/30.7.1575.

Enright, Anton J, and Christos A Ouzounis. 2001. “Functional Associations of Proteins in Entire Genomes by Means of Exhaustive Detection of Gene Fusions.” *Genome Biology* 2 (9). https://doi.org/10.1186/gb-2001-2-9-research0034.

Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. “Deep Learning: New Computational Modelling Techniques for Genomics.” *Nature Reviews Genetics* 20 (7): 389–403. https://doi.org/10.1038/s41576-019-0122-6.

Erdős, Gábor, Mátyás Pajkos, and Zsuzsanna Dosztányi. 2021. “IUPred3: Prediction of Protein Disorder Enhanced with Unambiguous Experimental Annotation and Visualization of Evolutionary Conservation.” *Nucleic Acids Research* 49 (W1): W297–303. https://doi.org/10.1093/nar/gkab408.

Ermel, Utz, Anchi Cheng, Jun Xi Ni, et al. 2024. “A Data Portal for Providing Standardized Annotations for Cryo-Electron Tomography.” *Nature Methods* 21 (12): 2200–2202. https://doi.org/10.1038/s41592-024-02477-2.

Evans, Richard, Michael O’Neill, Alexander Pritzel, et al. 2021. “Protein Complex Prediction with AlphaFold-Multimer.” *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2021.10.04.463034.

EvolutionaryScale. 2024. *ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning*. https://www.evolutionaryscale.ai/blog/esm-cambrian.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Ewing, Brent, LaDeana Hillier, Michael C. Wendl, and Phil Green. 1998. “Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment.” *Genome Research* 8 (3): 175–85. https://doi.org/10.1101/gr.8.3.175.

Ezkurdia, Iakes, David Juan, Jose Manuel Rodriguez, et al. 2014. “Multiple Evidence Strands Suggest That There May Be as Few as 19 000 Human Protein-Coding Genes.” *Human Molecular Genetics* 23 (22): 5866–78. https://doi.org/10.1093/hmg/ddu309.

Ezkurdia, Iakes, Jesús Vázquez, Alfonso Valencia, and Michael Tress. 2014. “Analyzing the First Drafts of the Human Proteome.” *Journal of Proteome Research* 13 (8): 3854–55. https://doi.org/10.1021/pr500572z.

Falk, Thorsten, Dominic Mai, Robert Bensch, et al. 2018. “U-Net: Deep Learning for Cell Counting, Detection, and Morphometry.” *Nature Methods* 16 (1): 67–70. https://doi.org/10.1038/s41592-018-0261-2.

Farrar, Michael. 2006. “Striped Smith–Waterman Speeds Database Searches Six Times over Other SIMD Implementations.” *Bioinformatics* 23 (2): 156–61. https://doi.org/10.1093/bioinformatics/btl582.

Fay, Marta M., Oren Kraus, Mason Victors, et al. 2023. “RxRx3: Phenomics Map of Biology.” *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2023.02.07.527350.

Felsenstein, J. 1978. "Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading." *Systematic Biology* 27 (4): 401–10. https://doi.org/10.1093/sysbio/27.4.401.

Felsenstein, Joseph. 1981. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *Journal of Molecular Evolution* 17 (6): 368–76. https://doi.org/10.1007/bf01734359.

Felsenstein, Joseph. 1985. "CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP." *Evolution* 39 (4): 783–91. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sinauer Associates.

Feng, Da-Fei, and Russell F. Doolittle. 1987. "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees." *Journal of Molecular Evolution* 25 (4): 351–60. https://doi.org/10.1007/bf02603120.

Feng, Haonan, Lang Wu, Bingxin Zhao, et al. 2025. "Benchmarking DNA Foundation Models for Genomic and Genetic Tasks." *Nature Communications* 16 (1). https://doi.org/10.1038/s41467-025-65823-8.

Fenn, John B., Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse. 1989. "Electrospray Ionization for Mass Spectrometry of Large Biomolecules." *Science* 246 (4926): 64–71. https://doi.org/10.1126/science.2675315.

Fernández, Pol, Rémy Amice, David Bruy, et al. 2024. "A 160 Gbp Fork Fern Genome Shatters Size Record for Eukaryotes." *iScience* 27 (6): 109889. https://doi.org/10.1016/j.isci.2024.109889.

Fernandez-Escamilla, Ana-Maria, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. 2004. "Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins." *Nature Biotechnology* 22 (10): 1302–6. https://doi.org/10.1038/nbt1012.

Ferragina, P., and G. Manzini. 2000. "Opportunistic Data Structures with Applications." *Proceedings 41st Annual Symposium on Foundations of Computer Science*, SFCS-00, 390–98. https://doi.org/10.1109/sfcs.2000.892127.

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker. 2022. "ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-32007-7.

Fickett, James W. 1982. "Recognition of Protein Coding Regions in DNA Sequences." *Nucleic Acids Research* 10 (17): 5303–18. https://doi.org/10.1093/nar/10.17.5303.

Fields, Stanley, and Ok-kyu Song. 1989. "A Novel Genetic System to Detect Protein–Protein Interactions." *Nature* 340 (6230): 245–46. https://doi.org/10.1038/340245a0.

Finn, Robert D., Jaina Mistry, Benjamin Schuster-Böckler, et al. 2006. "Pfam: Clans, Web Tools and Services." *Nucleic Acids Research* 34 (90001): D247–51. https://doi.org/10.1093/nar/gkj149.

Fire, Andrew, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. 1998. "Potent and Specific Genetic Interference by Double-Stranded RNA in Caenorhabditis Elegans." *Nature* 391 (6669): 806–11. https://doi.org/10.1038/35888.

Fischer, Daniel, and David Eisenberg. 1996. "Protein Fold Recognition Using Sequence-Derived Predic-

tions." *Protein Science* 5 (5): 947–55. https://doi.org/10.1002/pro.5560050516.

Fischer, Samantha N, Erin R Claussen, Savvas Kourtis, et al. 2025. "Hu.MAP3.0: Atlas of Human Protein Complexes by Integration of >25,000 Proteomic Experiments." *Molecular Systems Biology* 21 (7). https://doi.org/10.1038/s44320-025-00121-5.

Fiser, András, Richard Kinh Gian Do, and Andrej Šali. 2000. "Modeling of Loops in Protein Structures." *Protein Science* 9 (9): 1753–73. https://doi.org/10.1110/ps.9.9.1753.

Fitch, Walter M. 1970. "Distinguishing Homologous from Analogous Proteins." *Systematic Zoology* 19 (2): 99. https://doi.org/10.2307/2412448.

Fitch, Walter M. 1971. "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology." *Systematic Zoology* 20 (4): 406. https://doi.org/10.2307/2412116.

Fitch, Walter M. 2000. "Homology." *Trends in Genetics* 16 (5): 227–31. https://doi.org/10.1016/s0168-9525(00)02005-9.

Fitch, Walter M., and Emanuel Margoliash. 1967. "Construction of Phylogenetic Trees: A Method Based on Mutation Distances as Estimated from Cytochrome c Sequences Is of General Applicability." *Science* 155 (3760): 279–84. https://doi.org/10.1126/science.155.3760.279.

Fitzpatrick, Anthony W. P., Benjamin Falcon, Shaoda He, et al. 2017. "Cryo-EM Structures of Tau Filaments from Alzheimer's Disease." *Nature* 547 (7662): 185–90. https://doi.org/10.1038/nature23002.

Fleischmann, Robert D., Mark D. Adams, Owen White, et al. 1995. "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd." *Science* 269 (5223): 496–512. https://doi.org/10.1126/science.7542800.

Fleming, Jennifer, Paulyna Magana, Sreenath Nair, et al. 2025. "AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities." *Journal of Molecular Biology* 437 (15): 168967. https://doi.org/10.1016/j.jmb.2025.168967.

Fodor, Stephen P. A., J. Leighton Read, Michael C. Pirrung, Lubert Stryer, Amy Tsai Lu, and Dennis Solas. 1991. "Light-Directed, Spatially Addressable Parallel Chemical Synthesis." *Science* 251 (4995): 767–73. https://doi.org/10.1126/science.1990438.

Force, Allan, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. 1999. "Preservation of Duplicate Genes by Complementary, Degenerative Mutations." *Genetics* 151 (4): 1531–45. https://doi.org/10.1093/genetics/151.4.1531.

Forslund, Sofia K., Mateusz Kaduk, and Erik L. L. Sonnhammer. 2019. "Evolution of Protein Domain Architectures." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_15.

Foster, Peter G. 2004. "Modeling Compositional Heterogeneity." *Systematic Biology* 53 (3): 485–95. https://doi.org/10.1080/10635150490445779.

Fowler, Douglas M, and Stanley Fields. 2014. "Deep Mutational Scanning: A New Style of Protein Science." *Nature Methods* 11 (8): 801–7. https://doi.org/10.1038/nmeth.3027.

Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. 2013. "SCOPe: Structural Classification of Proteins—Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures." *Nucleic Acids Research* 42 (D1): D304–9. https://doi.org/10.1093/nar/gkt1240.

Fraser, Claire M., Jeannine D. Gocayne, Owen White, et al. 1995. "The Minimal Gene Complement of Mycoplasma Genitalium." *Science* 270 (5235): 397–404. https://doi.org/10.1126/science.270.5235.397.

Frazer, Jonathan, Pascal Notin, Mafalda Dias, et al. 2021. "Disease Variant Prediction with Deep Generative Models of Evolutionary Data." *Nature* 599 (7883): 91–95. https://doi.org/10.1038/s41586-021-04043-8.

Friedberg, Iddo, and Predrag Radivojac. 2016. "Community-Wide Evaluation of Computational Function Prediction." In *The Gene Ontology Handbook*. Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_10.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5). https://doi.org/10.1214/aos/1013203451.

Frishman, Dmitrij, and Patrick Argos. 1995. "Knowledge-Based Protein Secondary Structure Assignment." *Proteins: Structure, Function, and Bioinformatics* 23 (4): 566–79. https://doi.org/10.1002/prot.340230412.

Frommer, M, L E McDonald, D S Millar, et al. 1992. "A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands." *Proceedings of the National Academy of Sciences* 89 (5): 1827–31. https://doi.org/10.1073/pnas.89.5.1827.

Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. 2016. "Formation of Chromosomal Domains by Loop Extrusion." *Cell Reports* 15 (9): 2038–49. https://doi.org/10.1016/j.celrep.2016.04.085.

Gabaldón, Toni, and Eugene V. Koonin. 2013. "Functional and Evolutionary Implications of Gene Orthology." *Nature Reviews Genetics* 14 (5): 360–66. https://doi.org/10.1038/nrg3456.

Gabriel, Lars, Tomáš Brůna, Katharina J. Hoff, et al. 2024. "BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA." *Genome Research* 34 (5): 769–77. https://doi.org/10.1101/gr.278090.123.

Gabriel, Stacey B., Stephen F. Schaffner, Huy Nguyen, et al. 2002. "The Structure of Haplotype Blocks in the Human Genome." *Science* 296 (5576): 2225–29. https://doi.org/10.1126/science.1069424.

Gabriels, Ralf, Lennart Martens, and Sven Degroeve. 2019. "Updated MS²PIP Web Server Delivers Fast and Accurate MS² Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques." *Nucleic Acids Research* 47 (W1): W295–99. https://doi.org/10.1093/nar/gkz299.

Gagniuc, Paul A. 2021. *Algorithms in Bioinformatics: Theory and Implementation*. Wiley. https://doi.org/10.1002/9781119698005.

Gao, Ruli, Shanshan Bai, Ying C. Henderson, et al. 2021. "Delineating Copy Number and Clonal Substructure in Human Tumors from Single-Cell Transcriptomes." *Nature Biotechnology* 39 (5): 599–608. https://doi.org/10.1038/s41587-020-00795-2.

Garalde, Daniel R, Elizabeth A Snell, Daniel Jachimowicz, et al. 2018. "Highly Parallel Direct RNA Sequencing on an Array of Nanopores." *Nature Methods* 15 (3): 201–6. https://doi.org/10.1038/nmeth.4577.

Gardner, Timothy S., Diego di Bernardo, David Lorenz, and James J. Collins. 2003. "Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling." *Science* 301 (5629): 102–5. https://doi.org/10.1126/science.1081900.

Gardner, Timothy S., Charles R. Cantor, and James J. Collins. 2000. "Construction of a Genetic Toggle Switch in Escherichia Coli." *Nature* 403 (6767): 339–42. https://doi.org/10.1038/35002131.

Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. "Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins." *Journal of Molecular Biology* 120 (1): 97–120. https://doi.org/10.1016/0022-2836(78)90297-8.

Garrido-Ramos, Manuel. 2017. "Satellite DNA: An Evolving Topic." *Genes* 8 (9): 230. https://doi.org/10.3390/genes8090230.

Garrison, Erik, Jouni Sirén, Adam M Novak, et al. 2018. "Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference." *Nature Biotechnology* 36 (9): 875–79. https://doi.org/10.1038/nbt.4227.

Garzón, José Ignacio, Lei Deng, Diana Murray, Sagi Shapira, Donald Petrey, and Barry Honig. 2016. "A Computational Interactome and Functional Annotation for the Human Proteome." *eLife* 5. https://doi.org/10.7554/elife.18715.

Gascuel, O. 2006. "Neighbor-Joining Revealed." *Molecular Biology and Evolution* 23 (11): 1997–2000. https://doi.org/10.1093/molbev/msl072.

Gaudet, Pascale, Nives Škunca, James C. Hu, and Christophe Dessimoz. 2016. "Primer on the Gene Ontology." In *The Gene Ontology Handbook*. Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_3.

Gaudet, P., M. S. Livstone, S. E. Lewis, and P. D. Thomas. 2011. "Phylogenetic-Based Propagation of Functional Annotations Within the Gene Ontology Consortium." *Briefings in Bioinformatics* 12 (5): 449–62. https://doi.org/10.1093/bib/bbr042.

Gavin, Anne-Claude, Markus Bösche, Roland Krause, et al. 2002. "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes." *Nature* 415 (6868): 141–47. https://doi.org/10.1038/415141a.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. "Single-Cell Genome Sequencing: Current State of the Science." *Nature Reviews Genetics* 17 (3): 175–88. https://doi.org/10.1038/nrg.2015.16.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31. https://doi.org/10.1198/000313006x152649.

Genereux, Diane P., Aitor Serres, Joel Armstrong, et al. 2020. "A Comparative Genomics Multitool for Scientific Discovery and Conservation." *Nature* 587 (7833): 240–45. https://doi.org/10.1038/s41586-020-2876-6.

Gerasimavicius, Lukas, Sarah A. Teichmann, and Joseph A. Marsh. 2025. "Leveraging Protein Structural Information to Improve Variant Effect Prediction." *Current Opinion in Structural Biology* 92: 103023. https://doi.org/10.1016/j.sbi.2025.103023.

Gerber, Scott A., John Rush, Olaf Stemman, Marc W. Kirschner, and Steven P. Gygi. 2003. "Absolute Quantification of Proteins and Phosphoproteins from Cell Lysates by Tandem MS." *Proceedings of the National Academy of Sciences* 100 (12): 6940–45. https://doi.org/10.1073/pnas.0832254100.

Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, et al. 2012. "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing." *New England Journal of Medicine* 366 (10): 883–92. https://doi.org/10.1056/nejmoa1113205.

Gerstein, Mark B., Can Bruce, Joel S. Rozowsky, et al. 2007. "What Is a Gene, Post-ENCODE? History and Updated Definition." *Genome Research* 17 (6): 669–81. https://doi.org/10.1101/gr.6339607.

Gessulat, Siegfried, Tobias Schmidt, Daniel Paul Zolg, et al. 2019. "Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning." *Nature Methods* 16 (6): 509–18. https://doi.org/10.1038/s41592-019-0426-7.

Giaever, Guri, Angela M. Chu, Li Ni, et al. 2002. "Functional Profiling of the Saccharomyces Cerevisiae Genome." *Nature* 418 (6896): 387–91. https://doi.org/10.1038/nature00935.

Giambartolomei, Claudia, Damjan Vukcevic, Eric E. Schadt, et al. 2014. "Bayesian Test for Colocalisation Between Pairs of Genetic Association Studies Using Summary Statistics." *PLoS Genetics* 10 (5): e1004383. https://doi.org/10.1371/journal.pgen.1004383.

Gianni, Stefano, María Inés Freiberger, Per Jemth, Diego U. Ferreiro, Peter G. Wolynes, and Monika Fuxreiter. 2021. "Fuzziness and Frustration in the Energy Landscape of Protein Folding, Function, and Assembly." *Accounts of Chemical Research* 54 (5): 1251–59. https://doi.org/10.1021/acs.accounts.0c00813.

Gibbons, Francis D., and Frederick P. Roth. 2002. "Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation." *Genome Research* 12 (10): 1574–81. https://doi.org/10.1101/gr.397002.

Gibbs, Adrian J., and George A. Mcintyre. 1970. "The Diagram, a Method for Comparing Sequences: Its Use with Amino Acid and Nucleotide Sequences." *European Journal of Biochemistry* 16 (1): 1–11. https://doi.org/10.1111/j.1432-1033.1970.tb01046.x.

Gibson, Greg, Joseph E Powell, and Urko M Marigorta. 2015. "Expression Quantitative Trait Locus Analysis for Translational Medicine." *Genome Medicine* 7 (1). https://doi.org/10.1186/s13073-015-0186-7.

Giesen, Charlotte, Hao A O Wang, Denis Schapiro, et al. 2014. "Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry." *Nature Methods* 11 (4): 417–22. https://doi.org/10.1038/nmeth.2869.

Gilbert, Luke A., Max A. Horlbeck, Britt Adamson, et al. 2014. "Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation." *Cell* 159 (3): 647–61. https://doi.org/10.1016/j.cell.2014.09.029.

Gilbert, Walter. 1986. "Origin of Life: The RNA World." *Nature* 319 (6055): 618–18. https://doi.org/10.1

038/319618a0.

Gilks, Walter R., Benjamin Audit, Daniela De Angelis, Sophia Tsoka, and Christos A. Ouzounis. 2002. "Modeling the Percolation of Annotation Errors in a Database of Protein Sequences." *Bioinformatics* 18 (12): 1641–49. https://doi.org/10.1093/bioinformatics/18.12.1641.

Gillani, Maryam, and Gianluca Pollastri. 2024. "Protein Subcellular Localization Prediction Tools <Sup/>." *Computational and Structural Biotechnology Journal* 23: 1796–807. https://doi.org/10.1016/j.csbj.2024.04.032.

Gillespie, Daniel T. 1977. "Exact Stochastic Simulation of Coupled Chemical Reactions." *The Journal of Physical Chemistry* 81 (25): 2340–61. https://doi.org/10.1021/j100540a008.

Gillet, Ludovic C., Pedro Navarro, Stephen Tate, et al. 2012. "Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis." *Molecular & Cellular Proteomics* 11 (6): O111.016717. https://doi.org/10.1074/mcp.o111.016717.

Gillis, Jesse, and Paul Pavlidis. 2011. "The Impact of Multifunctional Genes on 'Guilt by Association' Analysis." *PLoS ONE* 6 (2): e17258. https://doi.org/10.1371/journal.pone.0017258.

Giovannoni, Stephen J., H. James Tripp, Scott Givan, et al. 2005. "Genome Streamlining in a Cosmopolitan Oceanic Bacterium." *Science* 309 (5738): 1242–45. https://doi.org/10.1126/science.1114057.

Girvan, M., and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99 (12): 7821–26. https://doi.org/10.1073/pnas.122653799.

Gligorijević, Vladimir, P. Douglas Renfrew, Tomasz Kosciolek, et al. 2021. "Structure-Based Protein Function Prediction Using Graph Convolutional Networks." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-23303-9.

Glover, Natasha, Christophe Dessimoz, Ingo Ebersberger, et al. 2019. "Advances and Applications in the Quest for Orthologs." *Molecular Biology and Evolution* 36 (10): 2157–64. https://doi.org/10.1093/molbev/msz150.

Gnirke, Andreas, Alexandre Melnikov, Jared Maguire, et al. 2009. "Solution Hybrid Selection with Ultra-Long Oligonucleotides for Massively Parallel Targeted Sequencing." *Nature Biotechnology* 27 (2): 182–89. https://doi.org/10.1038/nbt.1523.

Göbel, Ulrike, Chris Sander, Reinhard Schneider, and Alfonso Valencia. 1994. "Correlated Mutations and Residue Contacts in Proteins." *Proteins: Structure, Function, and Bioinformatics* 18 (4): 309–17. https://doi.org/10.1002/prot.340180402.

Goffeau, A., B. G. Barrell, H. Bussey, et al. 1996. "Life with 6000 Genes." *Science* 274 (5287): 546–67. https://doi.org/10.1126/science.274.5287.546.

Gogarten, J P, H Kibak, P Dittrich, et al. 1989. "Evolution of the Vacuolar h+-ATPase: Implications for the Origin of Eukaryotes." *Proceedings of the National Academy of Sciences* 86 (17): 6661–65. https://doi.org/10.1073/pnas.86.17.6661.

Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. "The Human Disease Network." *Proceedings of the National Academy of Sciences* 104 (21): 8685–90. https://doi.org/10.1073/pnas.0701361104.

Gold, Larry, Deborah Ayers, Jennifer Bertino, et al. 2010. "Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery." *PLoS ONE* 5 (12): e15004. https://doi.org/10.1371/journal.pone.0015004.

Goldman, Nick, and Ziheng Yang. 1994. "A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences." *Molecular Biology and Evolution*, ahead of print. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Goltsev, Yury, Nikolay Samusik, Julia Kennedy-Darling, et al. 2018. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging." *Cell* 174 (4): 968–981.e15. https://doi.org/10.1016/j.cell.2018.07.010.

Golub, T. R., D. K. Slonim, P. Tamayo, et al. 1999. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science* 286 (5439): 531–37. https://doi.org/10.1126/science.286.5439.531.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63 (11): 139–44. https://doi.org/10.1145/3422622.

Goodman, Morris, John Czelusniak, G. William Moore, A. E. Romero-Herrera, and Genji Matsuda. 1979. "Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences." *Systematic Zoology* 28: 132–63.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of Next-Generation Sequencing Technologies." *Nature Reviews Genetics* 17 (6): 333–51. https://doi.org/10.1038/nrg.2016.49.

Gotea, V. 2003. "Mastering Seeds for Genomic Size Nucleotide BLAST Searches." *Nucleic Acids Research* 31 (23): 6935–41. https://doi.org/10.1093/nar/gkg886.

Gotoh, Osamu. 1982. "An Improved Algorithm for Matching Biological Sequences." *Journal of Molecular Biology* 162 (3): 705–8. https://doi.org/10.1016/0022-2836(82)90398-9.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data Without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Graur, D., Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, and E. Elhaik. 2013. "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE." *Genome Biology and Evolution* 5 (3): 578–90. https://doi.org/10.1093/gbe/evt028.

Gray, Jeffrey J., Stewart Moughon, Chu Wang, et al. 2003. "Protein–Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations." *Journal of Molecular Biology* 331 (1): 281–99. https://doi.org/10.1016/s0022-2836(03)00670-3.

Green, Anna G., Hadeer Elhabashy, Kelly P. Brock, Rohan Maddamsetti, Oliver Kohlbacher, and Debora S. Marks. 2021. "Large-Scale Discovery of Protein Interactions at Residue Resolution Using Co-

Evolution Calculated from Genomic Sequences." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-21636-z.

Green, Richard E., Johannes Krause, Adrian W. Briggs, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Greener, Joe G., Shaun M. Kandathil, Lewis Moffat, and David T. Jones. 2021. "A Guide to Machine Learning for Biologists." *Nature Reviews Molecular Cell Biology* 23 (1): 40–55. https://doi.org/10.1038/s41580-021-00407-0.

Greenwald, Noah F., Geneva Miller, Erick Moen, et al. 2021. "Whole-Cell Segmentation of Tissue Images with Human-Level Performance Using Large-Scale Data Annotation and Deep Learning." *Nature Biotechnology* 40 (4): 555–65. https://doi.org/10.1038/s41587-021-01094-0.

Gregory, T. Ryan. 2001. "Coincidence, Coevolution, or Causation? DNA Content, Cell Size, and the c-Value Enigma." *Biological Reviews of the Cambridge Philosophical Society* 76 (1): 65–101. https://doi.org/10.1017/s1464793100005595.

Gregory, T. Ryan. 2008. "Understanding Evolutionary Trees." *Evolution: Education and Outreach* 1 (2): 121–37. https://doi.org/10.1007/s12052-008-0035-x.

Greider, Carol W., and Elizabeth H. Blackburn. 1985. "Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts." *Cell* 43 (2): 405–13. https://doi.org/10.1016/0092-8674(85)90170-9.

Gremer, Lothar, Daniel Schölzel, Carla Schenk, et al. 2017. "Fibril Structure of Amyloid-β(1–42) by Cryo–Electron Microscopy." *Science* 358 (6359): 116–19. https://doi.org/10.1126/science.aao2825.

Gribskov, Michael, and Nina L. Robinson. 1996. "Use of Receiver Operating Characteristic (ROC) Analysis to Evaluate Sequence Matching." *Computers & Chemistry* 20 (1): 25–33. https://doi.org/10.1016/s0097-8485(96)80004-0.

Gribskov, M, A D McLachlan, and D Eisenberg. 1987. "Profile Analysis: Detection of Distantly Related Proteins." *Proceedings of the National Academy of Sciences* 84 (13): 4355–58. https://doi.org/10.1073/pnas.84.13.4355.

Griffiths-Jones, Sam, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. 2003. "Rfam: An RNA Family Database." *Nucleic Acids Research* 31 (1): 439–41. https://doi.org/10.1093/nar/gkg006.

Grimm, Dominik G., Chloé-Agathe Azencott, Fabian Aicheler, et al. 2015. "The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity." *Human Mutation* 36 (5): 513–23. https://doi.org/10.1002/humu.22768.

Grossmann, Steffen, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. 2007. "Improved Detection of Overrepresentation of Gene-Ontology Annotations with Parent–Child Analysis." *Bioinformatics* 23 (22): 3024–31. https://doi.org/10.1093/bioinformatics/btm440.

Gu, Changdai, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. 2019. "Current Status and Applications of Genome-Scale Metabolic Models." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1730-3.

Guala, Dimitri, Christoph Ogris, Nikola Müller, and Erik L L Sonnhammer. 2019. "Genome-Wide Functional Association Networks: Background, Data & State-of-the-Art Resources." *Briefings in Bioinformatics* 21 (4): 1224–37. https://doi.org/10.1093/bib/bbz064.

Guala, Dimitri, Erik Sjölund, and Erik L. L. Sonnhammer. 2014. "MaxLink: Network-Based Prioritization of Genes Tightly Linked to a Disease Seed Set." *Bioinformatics* 30 (18): 2689–90. https://doi.org/10.1093/bioinformatics/btu344.

Guerrero-Ferreira, Ricardo, Nicholas MI Taylor, Daniel Mona, et al. 2018. "Cryo-EM Structure of Alpha-Synuclein Fibrils." *eLife* 7. https://doi.org/10.7554/elife.36402.

Guerrier-Takada, Cecilia, Katheleen Gardiner, Terry Marsh, Norman Pace, and Sidney Altman. 1983. "The RNA Moiety of Ribonuclease p Is the Catalytic Subunit of the Enzyme." *Cell* 35 (3): 849–57. https://doi.org/10.1016/0092-8674(83)90117-4.

Guigó, Roderic, Steen Knudsen, Neil Drake, and Temple Smith. 1992. "Prediction of Gene Structure." *Journal of Molecular Biology* 226 (1): 141–57. https://doi.org/10.1016/0022-2836(92)90130-c.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21. https://doi.org/10.1093/sysbio/syq010.

Gunderson, Kevin L., Semyon Kruglyak, Michael S. Graige, et al. 2004. "Decoding Randomly Ordered DNA Arrays." *Genome Research* 14 (5): 870–77. https://doi.org/10.1101/gr.2255804.

Guo, Fei, Renchu Guan, Yaohang Li, et al. 2025. "Foundation Models in Bioinformatics." *National Science Review* 12 (4). https://doi.org/10.1093/nsr/nwaf028.

Gupta, Ankit, Zoe Wefers, Konstantin Kahnert, et al. 2024. "SubCell: Proteome-Aware Vision Foundation Models for Microscopy Capture Single-Cell Biology." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2024.12.06.627299.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Gustafsdottir, Sigrun M., Vebjorn Ljosa, Katherine L. Sokolnicki, et al. 2013. "Multiplex Cytological Profiling Assay to Measure Diverse Cellular States." *PLoS ONE* 8 (12): e80999. https://doi.org/10.1371/journal.pone.0080999.

Gutenkunst, Ryan N, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. 2007. "Universally Sloppy Parameter Sensitivities in Systems Biology Models." *PLoS Computational Biology* 3 (10): e189. https://doi.org/10.1371/journal.pcbi.0030189.

Haas, Jürgen, Alessandro Barbato, Dario Behringer, et al. 2017. "Continuous Automated Model EvaluatiOn (CAMEO) Complementing the Critical Assessment of Structure Prediction in CASP12." *Proteins: Structure, Function, and Bioinformatics* 86 (S1): 387–98. https://doi.org/10.1002/prot.25431.

Hafemeister, Christoph, and Rahul Satija. 2019. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." *Genome Biology* 20 (1).

https://doi.org/10.1186/s13059-019-1874-1.

Hahn, Matthew W., Tijl De Bie, Jason E. Stajich, Chi Nguyen, and Nello Cristianini. 2005. "Estimating the Tempo and Mode of Gene Family Evolution from Comparative Genomic Data." *Genome Research* 15 (8): 1153–60. https://doi.org/10.1101/gr.3567505.

Hakes, Luke, John W Pinney, David L Robertson, and Simon C Lovell. 2008. "Protein-Protein Interaction Networks and Biology—What's the Connection?" *Nature Biotechnology* 26 (1): 69–72. https://doi.org/10.1038/nbt0108-69.

Hallgren, Jeppe, Konstantinos D. Tsirigos, Mads Damgaard Pedersen, et al. 2022. "DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2022.04.08.487609.

Hallou, Adrien, Hannah G. Yevick, Bianca Dumitrascu, and Virginie Uhlmann. 2021. "Deep Learning for Bioimage Analysis in Developmental Biology." *Development* 148 (18). https://doi.org/10.1242/dev.199616.

Hamamsy, Tymor, James T. Morton, Robert Blackwell, et al. 2023. "Protein Remote Homology Detection and Structural Alignment Using Deep Learning." *Nature Biotechnology* 42 (6): 975–85. https://doi.org/10.1038/s41587-023-01917-2.

Han, Jing-Dong J., Nicolas Bertin, Tong Hao, et al. 2004. "Evidence for Dynamically Organized Modularity in the Yeast Protein–Protein Interaction Network." *Nature* 430 (6995): 88–93. https://doi.org/10.1038/nature02555.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. 1998. "Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products." *Chemistry & Biology* 5 (10): R245–49. https://doi.org/10.1016/s1074-5521(98)90108-9.

Hannenhalli, Sridhar, and Pavel A. Pevzner. 1999. "Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals." *Journal of the ACM* 46 (1): 1–27. https://doi.org/10.1145/300515.300516.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Haque, Ashraful, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. 2017. "A Practical Guide to Single-Cell RNA-Sequencing for Biomedical Research and Clinical Applications." *Genome Medicine* 9 (1). https://doi.org/10.1186/s13073-017-0467-4.

Hart, Traver, Megha Chandrashekhar, Michael Aregger, et al. 2015. "High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities." *Cell* 163 (6): 1515–26. https://doi.org/10.1016/j.cell.2015.11.015.

Hartley, Matthew, Gerard J. Kleywegt, Ardan Patwardhan, Ugis Sarkans, Jason R. Swedlow, and Alvis Brazma. 2022. "The BioImage Archive – Building a Home for Life-Sciences Microscopy Data." *Journal of Molecular Biology* 434 (11): 167505. https://doi.org/10.1016/j.jmb.2022.167505.

Hartwell, Leland H., John J. Hopfield, Stanislas Leibler, and Andrew W. Murray. 1999. "From Molecular to Modular Cell Biology." *Nature* 402 (S6761): C47–52. https://doi.org/10.1038/35011540.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA." *Journal of Molecular Evolution* 22 (2): 160–74. https://doi.org/10.1007/bf02101694.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. 2017. "Multi-Omics Approaches to Disease." *Genome Biology* 18 (1). https://doi.org/10.1186/s13059-017-1215-1.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer. https://hastie.su.domains/ElemStatLearn/.

Hastings, P. J., Grzegorz Ira, and James R. Lupski. 2009. "A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation." *PLoS Genetics* 5 (1): e1000327. https://doi.org/10.1371/journal.pgen.1000327.

Hastings, P. J., James R. Lupski, Susan M. Rosenberg, and Grzegorz Ira. 2009. "Mechanisms of Change in Gene Copy Number." *Nature Reviews Genetics* 10 (8): 551–64. https://doi.org/10.1038/nrg2593.

Havugimana, Pierre C., G. Traver Hart, Tamás Nepusz, et al. 2012. "A Census of Human Soluble Protein Complexes." *Cell* 150 (5): 1068–81. https://doi.org/10.1016/j.cell.2012.08.011.

Hayat, Sikander, Christoph Peters, Nanjiang Shu, Konstantinos D. Tsirigos, and Arne Elofsson. 2016. "Inclusion of Dyad-Repeat Pattern Improves Topology Prediction of Transmembrane β-Barrel Proteins." *Bioinformatics* 32 (10): 1571–73. https://doi.org/10.1093/bioinformatics/btw025.

Hayes, Thomas, Roshan Rao, Halil Akin, et al. 2025. "Simulating 500 Million Years of Evolution with a Language Model." *Science* 387 (6736): 850–58. https://doi.org/10.1126/science.ads0018.

He, Bryan, Ludvig Bergenstråhle, Linnea Stenbeck, et al. 2020. "Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning." *Nature Biomedical Engineering* 4 (8): 827–34. https://doi.org/10.1038/s41551-020-0578-x.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. 2022. "Masked Autoencoders Are Scalable Vision Learners." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–88. https://doi.org/10.1109/cvpr52688.2022.01553.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–78. https://doi.org/10.1109/cvpr.2016.90.

He, Xionglei, and Jianzhi Zhang. 2006. "Why Do Hubs Tend to Be Essential in Protein Networks?" *PLoS Genetics* 2 (6): e88. https://doi.org/10.1371/journal.pgen.0020088.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Hedlund, Brian P., Maria Chuvochina, Philip Hugenholtz, et al. 2022. "SeqCode: A Nomenclatural Code for Prokaryotes Described from Sequence Data." *Nature Microbiology*, ahead of print. https://doi.org/10.1038/s41564-022-01214-9.

Heijne, Gunnar von. 1986. "A New Method for Predicting Signal Sequence Cleavage Sites." *Nucleic Acids Research* 14 (11): 4683–90. https://doi.org/10.1093/nar/14.11.4683.

Heijne, Gunnar von. 1992. "Membrane Protein Structure Prediction." *Journal of Molecular Biology* 225 (2): 487–94. https://doi.org/10.1016/0022-2836(92)90934-c.

Heinz, Sven, Christopher Benner, Nathanael Spann, et al. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and b Cell Identities." *Molecular Cell* 38 (4): 576–89. https://doi.org/10.1016/j.molcel.2010.05.004.

Heinzinger, Michael, Ahmed Elnaggar, Yu Wang, et al. 2019. "Modeling Aspects of the Language of Life Through Transfer-Learning Protein Sequences." *BMC Bioinformatics* 20 (1). https://doi.org/10.1186/s12859-019-3220-8.

Heinzinger, Michael, Konstantin Weissenow, Joaquin Gomez Sanchez, et al. 2024. "Bilingual Language Model for Protein Sequence and Structure." *NAR Genomics and Bioinformatics* 6 (4). https://doi.org/10.1093/nargab/lqae150.

Helmy, Mohamed, Lingling Jin, Amr Alhossary, Tamer Mansour, Diogo Pellagrina, and Kumar Selvarajoo. 2025. "Ten Simple Rules for Optimal and Careful Use of Generative AI in Science." *PLOS Computational Biology* 21 (10): e1013588. https://doi.org/10.1371/journal.pcbi.1013588.

Henderson, R., J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing. 1990. "Model for the Structure of Bacteriorhodopsin Based on High-Resolution Electron Cryo-Microscopy." *Journal of Molecular Biology* 213 (4): 899–929. https://doi.org/10.1016/s0022-2836(05)80271-2.

Henikoff, S, and J G Henikoff. 1992. "Amino Acid Substitution Matrices from Protein Blocks." *Proceedings of the National Academy of Sciences* 89 (22): 10915–19. https://doi.org/10.1073/pnas.89.22.10915.

Henikoff, Steven, and Jorja G. Henikoff. 1994. "Position-Based Sequence Weights." *Journal of Molecular Biology* 243 (4): 574–78. https://doi.org/10.1016/0022-2836(94)90032-9.

Hennig, Willi. 1965. "Phylogenetic Systematics." *Annual Review of Entomology* 10 (1): 97–116. https://doi.org/10.1146/annurev.en.10.010165.000525.

Henzel, W J, T M Billeci, J T Stults, S C Wong, C Grimley, and C Watanabe. 1993. "Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases." *Proceedings of the National Academy of Sciences* 90 (11): 5011–15. https://doi.org/10.1073/pnas.90.11.5011.

Heumos, Lukas, Anna C. Schaar, Christopher Lance, et al. 2023. "Best Practices for Single-Cell Analysis Across Modalities." *Nature Reviews Genetics* 24 (8): 550–72. https://doi.org/10.1038/s41576-023-00586-w.

Higgins, Desmond G., and Paul M. Sharp. 1988. "CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer." *Gene* 73 (1): 237–44. https://doi.org/10.1016/0378-1119(88)90330-7.

Higgs, Paul G., and Niles Lehman. 2014. "The RNA World: Molecular Cooperation at the Origins of Life." *Nature Reviews Genetics* 16 (1): 7–17. https://doi.org/10.1038/nrg3841.

Hillenmeyer, Maureen E., Eula Fung, Jan Wildenhain, et al. 2008. "The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes." *Science* 320 (5874): 362–65. https://doi.org/10.1126/science.1150021.

Himmelstein, Daniel Scott, Antoine Lizee, Christine Hessler, et al. 2017. "Systematic Integration of Biomedical Knowledge Prioritizes Drugs for Repurposing." *eLife* 6. https://doi.org/10.7554/elife.26726.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. "Denoising Diffusion Probabilistic Models." *Advances in Neural Information Processing Systems* 33: 6840–51. https://arxiv.org/abs/2006.11239.

Ho, Simon Y. W., Matthew J. Phillips, Alan Cooper, and Alexei J. Drummond. 2005. "Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times." *Molecular Biology and Evolution* 22 (7): 1561–68. https://doi.org/10.1093/molbev/msi145.

Ho, Yuen, Albrecht Gruhler, Adrian Heilbut, et al. 2002. "Systematic Identification of Protein Complexes in Saccharomyces Cerevisiae by Mass Spectrometry." *Nature* 415 (6868): 180–83. https://doi.org/10.1038/415180a.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2017. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular Biology and Evolution* 35 (2): 518–22. https://doi.org/10.1093/molbev/msx281.

Hobohm, Uwe, Michael Scharf, Reinhard Schneider, and Chris Sander. 1992. "Selection of Representative Protein Data Sets." *Protein Science* 1 (3): 409–17. https://doi.org/10.1002/pro.5560010313.

Hobolth, Asger, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. 2007. "Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model." *PLoS Genetics* 3 (2): e7. https://doi.org/10.1371/journal.pgen.0030007.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Hofacker, Ivo L., Martin Fekete, and Peter F. Stadler. 2002. "Secondary Structure Prediction for Aligned RNA Sequences." *Journal of Molecular Biology* 319 (5): 1059–66. https://doi.org/10.1016/s0022-2836(02)00308-x.

Holley, L H, and M Karplus. 1989. "Protein Secondary Structure Prediction with a Neural Network." *Proceedings of the National Academy of Sciences* 86 (1): 152–56. https://doi.org/10.1073/pnas.86.1.152.

Holm, Liisa. 2019. "DALI and the Persistence of Protein Shape." *Protein Science* 29 (1): 128–40. https://doi.org/10.1002/pro.3749.

Holm, Liisa, and Chris Sander. 1993. "Protein Structure Comparison by Alignment of Distance Matrices." *Journal of Molecular Biology* 233 (1): 123–38. https://doi.org/10.1006/jmbi.1993.1489.

Holm, Liisa, and Chris Sander. 1996. "Mapping the Protein Universe." *Science* 273 (5275): 595–602. https://doi.org/10.1126/science.273.5275.595.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.

Holt, Carson, and Mark Yandell. 2011. "MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects." *BMC Bioinformatics* 12 (1). https://doi.org/10.1186/1471-2105-12-491.

Hong, Liang, Zhihang Hu, Siqi Sun, et al. 2024. "Fast, Sensitive Detection of Protein Homologs Using Deep Dense Retrieval." *Nature Biotechnology* 43 (6): 983–95. https://doi.org/10.1038/s41587-024-02353-6.

Hopf, Thomas A, John B Ingraham, Frank J Poelwijk, et al. 2017. "Mutation Effects Predicted from Sequence Co-Variation." *Nature Biotechnology* 35 (2): 128–35. https://doi.org/10.1038/nbt.3769.

Hopf, Thomas A, Charlotta P I Schärfe, João P G L M Rodrigues, et al. 2014. "Sequence Co-Evolution Gives 3D Contacts and Structures of Protein Complexes." *eLife* 3. https://doi.org/10.7554/elife.03430.

Hornbeck, Peter V., Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. 2014. "PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations." *Nucleic Acids Research* 43 (D1): D512–20. https://doi.org/10.1093/nar/gku1267.

Horne, Jim E., David J. Brockwell, and Sheena E. Radford. 2020. "Role of the Lipid Bilayer in Outer Membrane Protein Folding in Gram-Negative Bacteria." *Journal of Biological Chemistry* 295 (30): 10340–67. https://doi.org/10.1074/jbc.rev120.011473.

Howard, Frederick M., James Dolezal, Sara Kochanny, et al. 2021. "The Impact of Site-Specific Digital Histology Signatures on Deep Learning Model Accuracy and Bias." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-24698-1.

Hu, Qizhi, Robert J. Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R. Graham Cooks. 2005. "The Orbitrap: A New Mass Spectrometer." *Journal of Mass Spectrometry* 40 (4): 430–43. https://doi.org/10.1002/jms.856.

Huang, Da Wei, Brad T Sherman, and Richard A Lempicki. 2008. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57. https://doi.org/10.1038/nprot.2008.211.

Huang, Po-Ssu, Scott E. Boyken, and David Baker. 2016. "The Coming of Age of de Novo Protein Design." *Nature* 537 (7620): 320–27. https://doi.org/10.1038/nature19946.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, et al. 2017. "New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens." *Nature* 546 (7657): 289–92. https://doi.org/10.1038/nature22336.

HuBMAP Consortium. 2019. "The Human Body at Cellular Resolution: The NIH Human Biomolecular Atlas Program." *Nature* 574 (7777): 187–92. https://doi.org/10.1038/s41586-019-1629-x.

Hucka, M., A. Finney, H. M. Sauro, et al. 2003. "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models." *Bioinformatics* 19 (4): 524–31. https://doi.org/10.1093/bioinformatics/btg015.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology." *Science* 294 (5550): 2310–14. https://doi.org/10.1126/science.1065889.

Huerta-Cepas, Jaime, Salvador Capella-Gutiérrez, Leszek P. Pryszcz, Marina Marcet-Houben, and Toni Gabaldón. 2013. "PhylomeDB V4: Zooming into the Plurality of Evolutionary Histories of a Genome." *Nucleic Acids Research* 42 (D1): D897–902. https://doi.org/10.1093/nar/gkt1177.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, et al. 2018. "eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Huerta-Sánchez, Emilia, Xin Jin, Asan, et al. 2014. "Altitude Adaptation in Tibetans Caused by Introgression of Denisovan-Like DNA." *Nature* 512 (7513): 194–97. https://doi.org/10.1038/nature13408.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, et al. 2016. "A New View of the Tree of Life." *Nature Microbiology* 1 (5). https://doi.org/10.1038/nmicrobiol.2016.48.

Humphreys, Ian R., Jimin Pei, Minkyung Baek, et al. 2021. "Computed Structures of Core Eukaryotic Protein Complexes." *Science* 374 (6573). https://doi.org/10.1126/science.abm4805.

Humphreys, Ian R., Jing Zhang, Minkyung Baek, et al. 2024. "Protein Interactions in Human Pathogens Revealed Through Deep Learning." *Nature Microbiology* 9 (10): 2642–52. https://doi.org/10.1038/s41564-024-01791-x.

Huson, Daniel H., and David Bryant. 2005. "Application of Phylogenetic Networks in Evolutionary Studies." *Molecular Biology and Evolution* 23 (2): 254–67. https://doi.org/10.1093/molbev/msj030.

Hutchings, Charlotte, Charlotte S. Dawson, Thomas Krueger, Kathryn S. Lilley, and Lisa M. Breckels. 2023. "A Bioconductor Workflow for Processing, Evaluating, and Interpreting Expression Proteomics Data." *F1000Research* 12: 1402. https://doi.org/10.12688/f1000research.139116.1.

Huttlin, Edward L., Raphael J. Bruckner, Jose Navarrete-Perea, et al. 2021. "Dual Proteome-Scale Networks Reveal Cell-Specific Remodeling of the Human Interactome." *Cell* 184 (11): 3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011.

Huttlin, Edward L., Lily Ting, Raphael J. Bruckner, et al. 2015. "The BioPlex Network: A Systematic Exploration of the Human Interactome." *Cell* 162 (2): 425–40. https://doi.org/10.1016/j.cell.2015.06.043.

Huynen, M. A., and E. van Nimwegen. 1998. "The Frequency Distribution of Gene Family Sizes in Complete Genomes." *Molecular Biology and Evolution* 15 (5): 583–89. https://doi.org/10.1093/oxfordjournals.molbev.a025959.

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. 2010. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods." *PLoS ONE* 5 (9): e12776. https://doi.org/10.1371/journal.pone.0012776.

Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (1). https://doi.org/10.1186/1471-2105-11-119.

Ibarra, Rafael U., Jeremy S. Edwards, and Bernhard O. Palsson. 2002. "Escherichia Coli k-12 Undergoes Adaptive Evolution to Achieve in Silico Predicted Optimal Growth." *Nature* 420 (6912): 186–89. https://doi.org/10.1038/nature01149.

Ideker, T. E., V. Thorsson, and R. M. Karp. 2000. "Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design." *Pacific Symposium on Biocomputing*, 305–16.

Ideker, Trey, Owen Ozier, Benno Schwikowski, and Andrew F. Siegel. 2002. "Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks." *Bioinformatics* 18 (suppl_1): S233–40. https://doi.org/10.1093/bioinformatics/18.suppl_1.s233.

Ideker, Trey, Vesteinn Thorsson, Jeffrey A. Ranish, et al. 2001. "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network." *Science* 292 (5518): 929–34. https://doi.org/10.1126/science.292.5518.929.

Idury, Ramana M., and Michael S. Waterman. 1995. "A New Algorithm for DNA Sequence Assembly." *Journal of Computational Biology* 2 (2): 291–306. https://doi.org/10.1089/cmb.1995.2.291.

Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, et al. 2020. "Isolation of an Archaeon at the Prokaryote–Eukaryote Interface." *Nature* 577 (7791): 519–25. https://doi.org/10.1038/s41586-019-1916-6.

Imai, Kenichiro, and Kenta Nakai. 2020. "Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins from Their Amino Acid Sequences." *Frontiers in Genetics* 11. https://doi.org/10.3389/fgene.2020.607812.

Ingolia, Nicholas T., Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. 2009. "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling." *Science* 324 (5924): 218–23. https://doi.org/10.1126/science.1168978.

Ingraham, John B., Max Baranov, Zak Costello, et al. 2023. "Illuminating Protein Space with a Programmable Generative Model." *Nature* 623 (7989): 1070–78. https://doi.org/10.1038/s41586-023-06728-8.

Innan, Hideki, and Fyodor Kondrashov. 2010. "The Evolution of Gene Duplications: Classifying and Distinguishing Between Models." *Nature Reviews Genetics* 11 (2): 97–108. https://doi.org/10.1038/nrg2689.

International Human Genome Sequencing Consortium. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

International Human Genome Sequencing Consortium, Eric S. Lander, Lauren M. Linton, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Irizarry, R. A. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics* 4 (2): 249–64. https://doi.org/10.1093/biostatistics/4.2.249.

Islam, Saiful, Amit Zeisel, Simon Joost, et al. 2013. "Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers." *Nature Methods* 11 (2): 163–66. https://doi.org/10.1038/nmeth.2772.

Isomorphic Labs Team. 2026. *Accurate Predictions of Novel Biomolecular Interactions with IsoDDE.*

Technical report. Isomorphic Labs. https://doi.org/10.5281/zenodo.19699685.

Istrate, Ana-Maria, Fausto Milletari, Fabrizio Castrotorres, et al. 2025. "Rbio1 - Training Scientific Reasoning LLMs with Biological World Models as Soft Verifiers." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.08.18.670981.

Ito, Takashi, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. 2001. "A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome." *Proceedings of the National Academy of Sciences* 98 (8): 4569–74. https://doi.org/10.1073/pnas.061034498.

Iwabe, N, K Kuma, M Hasegawa, S Osawa, and T Miyata. 1989. "Evolutionary Relationship of Archaebacteria, Eubacteria, and Eukaryotes Inferred from Phylogenetic Trees of Duplicated Genes." *Proceedings of the National Academy of Sciences* 86 (23): 9355–59. https://doi.org/10.1073/pnas.86.23.9355.

Jackson, Aimee L, Steven R Bartz, Janell Schelter, et al. 2003. "Expression Profiling Reveals Off-Target Gene Regulation by RNAi." *Nature Biotechnology* 21 (6): 635–37. https://doi.org/10.1038/nbt831.

Jackson, Sophie E., and Alan R. Fersht. 1991. "Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition." *Biochemistry* 30 (43): 10428–35. https://doi.org/10.1021/bi00107a010.

Jacob, François, and Jacques Monod. 1961. "Genetic Regulatory Mechanisms in the Synthesis of Proteins." *Journal of Molecular Biology* 3 (3): 318–56. https://doi.org/10.1016/s0022-2836(61)80072-7.

Jacques, Florian, Paulina Bolivar, Kristian Pietras, and Emma U. Hammarlund. 2023. "Roadmap to the Study of Gene and Protein Phylogeny and Evolution—a Practical Guide." *PLOS ONE* 18 (2): e0279597. https://doi.org/10.1371/journal.pone.0279597.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-07641-9.

Jain, Miten, Sergey Koren, Karen H Miga, et al. 2018. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *Nature Biotechnology* 36 (4): 338–45. https://doi.org/10.1038/nbt.4060.

Jain, Sanjay, Liming Pei, Jeffrey M. Spraggins, et al. 2023. "Advances and Prospects for the Human BioMolecular Atlas Program (HuBMAP)." *Nature Cell Biology* 25 (8): 1089–100. https://doi.org/10.1038/s41556-023-01194-w.

Jakubowski & Flatt. 2024. *Amino Acids, Peptides, and Proteins (Chapter 3 of Fundamentals of Biochemistry)*. LibreTexts Biology (open textbook). https://bio.libretexts.org/Bookshelves/Biochemistry/Fundamentals_of_Biochemistry_(Jakubowski_and_Flatt)/01%3A_Unit_I-_Structure_and_Catalysis/03%3A_Amino_Acids_Peptides_and_Proteins.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. 2nd ed. Springer. https://www.statlearning.com/.

Jänes, Jürgen, and Pedro Beltrao. 2024. "Deep Learning for Protein Structure Prediction and Design—Progress and Applications." *Molecular Systems Biology* 20 (3): 162–69. https://doi.org/10.1038/s44320-024-00016-x.

Janesick, Amanda, Robert Shelansky, Andrew D. Gottscho, et al. 2023. "High Resolution Mapping of the Tumor Microenvironment Using Integrated Single-Cell, Spatial and in Situ Analysis." *Nature Communications* 14 (1). https://doi.org/10.1038/s41467-023-43458-x.

Janin, Joël, Kim Henrick, John Moult, et al. 2003. "CAPRI: A Critical Assessment of PRedicted Interactions." *Proteins: Structure, Function, and Bioinformatics* 52 (1): 2–9. https://doi.org/10.1002/prot.10381.

Jansen, Ronald, Haiyuan Yu, Dov Greenbaum, et al. 2003. "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data." *Science* 302 (5644): 449–53. https://doi.org/10.1126/science.1087361.

Jarmoskaite, Inga, Ishraq AlSadhan, Pavanapuresan P Vaidyanathan, and Daniel Herschlag. 2020. "How to Measure and Evaluate Binding Affinities." *eLife* 9. https://doi.org/10.7554/elife.57264.

Jeffroy, Olivier, Henner Brinkmann, Frédéric Delsuc, and Hervé Philippe. 2006. "Phylogenomics: The Beginning of Incongruence?" *Trends in Genetics* 22 (4): 225–31. https://doi.org/10.1016/j.tig.2006.02.003.

Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001. "Lethality and Centrality in Protein Networks." *Nature* 411 (6833): 41–42. https://doi.org/10.1038/35075138.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. 2000. "The Large-Scale Organization of Metabolic Networks." *Nature* 407 (6804): 651–54. https://doi.org/10.1038/35036627.

Ji, Yanrong, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. "DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome." *Bioinformatics* 37 (15): 2112–20. https://doi.org/10.1093/bioinformatics/btab083.

Jiang, Lin, Eric A. Althoff, Fernando R. Clemente, et al. 2008. "De Novo Computational Design of Retro-Aldol Enzymes." *Science* 319 (5868): 1387–91. https://doi.org/10.1126/science.1152692.

Jiang, Yuming, Devasahayam Arokia Balaya Rex, Dina Schuster, et al. 2024. "Comprehensive Overview of Bottom-up Proteomics Using Mass Spectrometry." *ACS Measurement Science Au* 4 (4): 338–417. https://doi.org/10.1021/acsmeasuresciau.3c00068.

Jiang, Yuxiang, Tal Ronnen Oron, Wyatt T. Clark, et al. 2016. "An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-1037-6.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 316 (5830): 1497–502. https://doi.org/10.1126/science.1141319.

Johnson, Jared L., Tomer M. Yaron, Emily M. Huntsman, et al. 2023. "An Atlas of Substrate Specificities for the Human Serine/Threonine Kinome." *Nature* 613 (7945): 759–66. https://doi.org/10.1038/s41586-022-05575-3.

Johnson, Jeff, Matthijs Douze, and Herve Jegou. 2021. "Billion-Scale Similarity Search with GPUs." *IEEE Transactions on Big Data* 7 (3): 535–47. https://doi.org/10.1109/tbdata.2019.2921572.

Jones, Christopher T., Edward Susko, and Joseph P. Bielawski. 2019. "Looking for Darwin in Ge-

nomic Sequences: Validity and Success Depends on the Relationship Between Model and Data." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_13.

Jones, D. T., W. R. Taylort, and J. M. Thornton. 1992. "A New Approach to Protein Fold Recognition." *Nature* 358 (6381): 86–89. https://doi.org/10.1038/358086a0.

Jones, David T. 1999. "Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices." *Journal of Molecular Biology* 292 (2): 195–202. https://doi.org/10.1006/jmbi.1999.3091.

Jones, David T. 2001. "Predicting Novel Protein Folds by Using FRAGFOLD." *Proteins: Structure, Function, and Genetics* 45 (S5): 127–32. https://doi.org/10.1002/prot.1171.

Jones, David T. 2019. "Setting the Standards for Machine Learning in Biology." *Nature Reviews Molecular Cell Biology* 20 (11): 659–60. https://doi.org/10.1038/s41580-019-0176-5.

Jones, David T., Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. 2011. "PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments." *Bioinformatics* 28 (2): 184–90. https://doi.org/10.1093/bioinformatics/btr638.

Jones, Neil C., and Pavel A. Pevzner. 2004. *An Introduction to Bioinformatics Algorithms*. MIT Press.

Jones, S, and J M Thornton. 1996. "Principles of Protein-Protein Interactions." *Proceedings of the National Academy of Sciences* 93 (1): 13–20. https://doi.org/10.1073/pnas.93.1.13.

Joy, Jeffrey B., Richard H. Liang, Rosemary M. McCloskey, T. Nguyen, and Art F. Y. Poon. 2016. "Ancestral Reconstruction." *PLOS Computational Biology* 12 (7): e1004763. https://doi.org/10.1371/journal.pcbi.1004763.

Joyce, Gerald F. 2002. "The Antiquity of RNA-Based Evolution." *Nature* 418 (6894): 214–21. https://doi.org/10.1038/418214a.

Jukes, Thomas H., and Charles R. Cantor. 1969. "Evolution of Protein Molecules." In *Mammalian Protein Metabolism*. Elsevier. https://doi.org/10.1016/b978-1-4832-3211-9.50009-7.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, David Silver, et al. 2021. "Applying and Improving AlphaFold at CASP14." *Proteins: Structure, Function, and Bioinformatics* 89 (12): 1711–21. https://doi.org/10.1002/prot.26257.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Jung, Hyungtaek, Tomer Ventura, J. Sook Chung, et al. 2020. "Twelve Quick Steps for Genome Assembly

and Annotation in the Classroom." *PLOS Computational Biology* 16 (11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325.

Kabsch, Wolfgang, and Christian Sander. 1983. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22 (12): 2577–637. https://doi.org/10.1002/bip.360221211.

Kaessmann, Henrik, Nicolas Vinckenbosch, and Manyuan Long. 2009. "RNA-Based Gene Duplication: Mechanistic and Evolutionary Insights." *Nature Reviews Genetics* 10 (1): 19–31. https://doi.org/10.1038/nrg2487.

Kajava, Andrey V. 2012. "Tandem Repeats in Proteins: From Sequence to Structure." *Journal of Structural Biology* 179 (3): 279–88. https://doi.org/10.1016/j.jsb.2011.08.009.

Käll, Lukas, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. 2007. "Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets." *Nature Methods* 4 (11): 923–25. https://doi.org/10.1038/nmeth1113.

Kalvari, Ioanna, Eric P Nawrocki, Nancy Ontiveros-Palacios, et al. 2020. "Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families." *Nucleic Acids Research* 49 (D1): D192–200. https://doi.org/10.1093/nar/gkaa1047.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." *Nature Methods* 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.

Kaminski, Kamil, Jan Ludwiczak, Kamil Pawlicki, Vikram Alva, and Stanislaw Dunin-Horkawicz. 2023. "pLM-BLAST: Distant Homology Detection Based on Direct Comparison of Sequence Representations from Protein Language Models." *Bioinformatics* 39 (10). https://doi.org/10.1093/bioinformatics/btad579.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. 2024. "KEGG: Biological Systems Database as a Model of the Real World." *Nucleic Acids Research* 53 (D1): D672–77. https://doi.org/10.1093/nar/gkae909.

Kang, Dongwan D., Feng Li, Edward Kirton, et al. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies." *PeerJ* 7: e7359. https://doi.org/10.7717/peerj.7359.

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. 2020. *Scaling Laws for Neural Language Models*. https://arxiv.org/abs/2001.08361.

Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. "Phylogenetic Tree Building in the Genomic Age." *Nature Reviews Genetics* 21 (7): 428–44. https://doi.org/10.1038/s41576-020-0233-0.

Kapoor, Sayash, and Arvind Narayanan. 2023. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science." *Patterns* 4 (9): 100804. https://doi.org/10.1016/j.patter.2023.100804.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte. 2017. "Dynamics of Genome Size Evolution in Birds and Mammals." *Proceedings of the National Academy of Sciences* 114 (8). https://doi.org/10.1073/pnas.1616702114.

Karas, Michael., and Franz. Hillenkamp. 1988. “Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons.” *Analytical Chemistry* 60 (20): 2299–301. https://doi.org/10.1021/ac00171a028.

Karlin, S, and S F Altschul. 1990. “Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes.” *Proceedings of the National Academy of Sciences* 87 (6): 2264–68. https://doi.org/10.1073/pnas.87.6.2264.

Karlin, S, and S F Altschul. 1993. “Applications and Statistics for Multiple High-Scoring Segments in Molecular Sequences.” *Proceedings of the National Academy of Sciences* 90 (12): 5873–77. https://doi.org/10.1073/pnas.90.12.5873.

Karlsson, Max, Cheng Zhang, Loren Méar, et al. 2021. “A Single–Cell Type Transcriptomics Map of Human Tissues.” *Science Advances* 7 (31). https://doi.org/10.1126/sciadv.abh2169.

Karplus, Martin, and J. Andrew McCammon. 2002. “Molecular Dynamics Simulations of Biomolecules.” *Nature Structural Biology* 9 (9): 646–52. https://doi.org/10.1038/nsb0902-646.

Karr, Jonathan R., Jayodita C. Sanghvi, Derek N. Macklin, et al. 2012. “A Whole-Cell Computational Model Predicts Phenotype from Genotype.” *Cell* 150 (2): 389–401. https://doi.org/10.1016/j.cell.2012.05.044.

Katchalski-Katzir, E, I Shariv, M Eisenstein, A A Friesem, C Aflalo, and I A Vakser. 1992. “Molecular Surface Recognition: Determination of Geometric Fit Between Proteins and Their Ligands by Correlation Techniques.” *Proceedings of the National Academy of Sciences* 89 (6): 2195–99. https://doi.org/10.1073/pnas.89.6.2195.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Katoh, K., and D. M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kedzierska, Kasia Z., Lorin Crawford, Ava P. Amini, and Alex X. Lu. 2025. “Zero-Shot Evaluation Reveals Limitations of Single-Cell Foundation Models.” *Genome Biology* 26 (1). https://doi.org/10.1186/s13059-025-03574-x.

Keller, Andrew, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. 2002. “Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.” *Analytical Chemistry* 74 (20): 5383–92. https://doi.org/10.1021/ac025747h.

Kelley, David R., Jasper Snoek, and John L. Rinn. 2016. “Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks.” *Genome Research* 26 (7): 990–99. https://doi.org/10.1101/gr.200535.115.

Kellis et al. 2016a. *Inferring Orthologs/Paralogs, Gene Duplication and Loss.* Computational Biology: Genomes, Networks, Evolution (MIT 6.047, LibreTexts). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/27%3A_Phylogenomics_II/27.05%3A_Inferring_Orthologs_Paralogs_Gene_Duplication_and_Loss.

Kellis et al. 2016b. *Natural Selection.* Computational Biology: Genomes, Networks, Evolution (MIT 6.047, LibreTexts). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/29%3A_Population_Genetic_Variation/29.04%3A_Natural_Selection.

Kellis et al. 2020a. *Hidden Markov Models i (Chapter 7 of Computational Biology: Genomes, Networks, and Evolution).* LibreTexts (MIT open textbook). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/07%3A_Hidden_Markov_Models_I.

Kellis et al. 2020b. *Rapid Sequence Alignment and Database Search (Chapter 3 of Computational Biology: Genomes, Networks, and Evolution).* LibreTexts (MIT open textbook). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/03%3A_Rapid_Sequence_Alignment_and_Database_Search.

Kellis et al. 2020c. *RNA Folding (Chapter 10 of Computational Biology: Genomes, Networks, and Evolution).* LibreTexts (MIT open textbook). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/10%3A_RNA_Folding.

Kellis et al. 2020d. *Sequence Alignment and Dynamic Programming (Chapter 2 of Computational Biology: Genomes, Networks, and Evolution).* LibreTexts (MIT open textbook). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/02%3A_Sequence_Alignment_and_Dynamic_Programming.

Kellis et al. 2021. *Molecular Evolution and Phylogenetics (Chapter 26 of Computational Biology: Genomes, Networks, and Evolution).* LibreTexts (MIT open textbook). https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/26%3A_Molecular_Evolution_and_Phylogenetics.

Kellis, Manolis, Bruce W. Birren, and Eric S. Lander. 2004. "Proof and Evolutionary Analysis of Ancient Genome Duplication in the Yeast Saccharomyces Cerevisiae." *Nature* 428 (6983): 617–24. https://doi.org/10.1038/nature02424.

Kellis, Manolis, Barbara Wold, Michael P. Snyder, et al. 2014. "Defining Functional DNA Elements in the Human Genome." *Proceedings of the National Academy of Sciences* 111 (17): 6131–38. https://doi.org/10.1073/pnas.1318948111.

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. 1958. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by x-Ray Analysis." *Nature* 181 (4610): 662–66. https://doi.org/10.1038/181662a0.

Kent, W. James. 2002. "BLAT —the BLAST -Like Alignment Tool." *Genome Research* 12 (4): 656–64. https://doi.org/10.1101/gr.229202.

Kent, W. James, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. 2003. "Evolution's Cauldron: Duplication, Deletion, and Rearrangement in the Mouse and Human Genomes." *Proceedings of the National Academy of Sciences* 100 (20): 11484–89. https://doi.org/10.1073/pnas.1932072100.

Kerfeld, Cheryl A., and Kathleen M. Scott. 2011. "Using BLAST to Teach 'e-Value-Tionary' Concepts." *PLoS Biology* 9 (2): e1001014. https://doi.org/10.1371/journal.pbio.1001014.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

Kim, Min-Sik, Sneha M. Pinto, Derese Getnet, et al. 2014. "A Draft Map of the Human Proteome." *Nature* 509 (7502): 575–81. https://doi.org/10.1038/nature13302.

Kim, Rachel Seongeun, Eli Levy Karin, Milot Mirdita, Rayan Chikhi, and Martin Steinegger. 2024. "BFVD—a Large Repository of Predicted Viral Protein Structures." *Nucleic Acids Research* 53 (D1): D340–47. https://doi.org/10.1093/nar/gkae1119.

Kim, Sangtae, and Pavel A. Pevzner. 2014. "MS-GF+ Makes Progress Towards a Universal Database Search Tool for Proteomics." *Nature Communications* 5 (1). https://doi.org/10.1038/ncomms6277.

Kim, Seung Joong, Javier Fernandez-Martinez, Ilona Nudelman, et al. 2018. "Integrative Structure and Functional Anatomy of a Nuclear Pore Complex." *Nature* 555 (7697): 475–82. https://doi.org/10.1038/nature26003.

Kimura, Motoo. 1968. "Evolutionary Rate at the Molecular Level." *Nature* 217 (5129): 624–26. https://doi.org/10.1038/217624a0.

Kimura, Motoo. 1980. "A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences." *Journal of Molecular Evolution* 16 (2): 111–20. https://doi.org/10.1007/bf01731581.

King, Jack Lester, and Thomas H. Jukes. 1969. "Non-Darwinian Evolution: Most Evolutionary Change in Proteins May Be Due to Neutral Mutations and Genetic Drift." *Science* 164 (3881): 788–98. https://doi.org/10.1126/science.164.3881.788.

Kingma, Diederik P., and Max Welling. 2014. *Auto-Encoding Variational Bayes*. https://arxiv.org/abs/1312.6114.

Kiouri, Despoina P., Georgios C. Batsis, and Christos T. Chasapis. 2025. "Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning." *Biomolecules* 15 (1): 141. https://doi.org/10.3390/biom15010141.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15. https://doi.org/10.1038/ng.2892.

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, et al. 2023. "Segment Anything." *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. https://doi.org/10.1109/iccv51070.2023.00371.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. "Optimization by Simulated Annealing." *Science* 220 (4598): 671–80. https://doi.org/10.1126/science.220.4598.671.

Kitano, Hiroaki. 2002. "Systems Biology: A Brief Overview." *Science* 295 (5560): 1662–64. https://doi.org/10.1126/science.1069492.

Kitano, Hiroaki. 2004. "Biological Robustness." *Nature Reviews Genetics* 5 (11): 826–37. https://doi.org/

10.1038/nrg1471.

Klein, Allon M., Linas Mazutis, Ilke Akartuna, et al. 2015. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells." *Cell* 161 (5): 1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

Knowles, David G., and Aoife McLysaght. 2009. "Recent de Novo Origin of Human Protein-Coding Genes." *Genome Research* 19 (10): 1752–59. https://doi.org/10.1101/gr.095026.109.

Kobak, Dmitry, and Philipp Berens. 2019. "The Art of Using t-SNE for Single-Cell Transcriptomics." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-13056-x.

Kobayashi, Hirofumi, Keith C. Cheveralls, Manuel D. Leonetti, and Loic A. Royer. 2022. "Self-Supervised Deep Learning Encodes High-Resolution Features of Protein Subcellular Localization." *Nature Methods* 19 (8): 995–1003. https://doi.org/10.1038/s41592-022-01541-z.

Koboldt, Daniel C. 2020. "Best Practices for Variant Calling in Clinical Sequencing." *Genome Medicine* 12 (1). https://doi.org/10.1186/s13073-020-00791-w.

Koga, Nobuyasu, Rie Tatsumi-Koga, Gaohua Liu, et al. 2012. "Principles for Designing Ideal Protein Structures." *Nature* 491 (7423): 222–27. https://doi.org/10.1038/nature11600.

Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. "Walking the Interactome for Prioritization of Candidate Disease Genes." *The American Journal of Human Genetics* 82 (4): 949–58. https://doi.org/10.1016/j.ajhg.2008.02.013.

Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, et al. 2020. "The Human Phenotype Ontology in 2021." *Nucleic Acids Research* 49 (D1): D1207–17. https://doi.org/10.1093/nar/gkaa1043.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Kolodny, Rachel, Patrice Koehl, and Michael Levitt. 2005. "Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures." *Journal of Molecular Biology* 346 (4): 1173–88. https://doi.org/10.1016/j.jmb.2004.12.032.

Kondo, T., S. Plaza, J. Zanet, et al. 2010. "Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis." *Science* 329 (5989): 336–39. https://doi.org/10.1126/science.1188158.

Kong, Andy T, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. 2017. "MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics." *Nature Methods* 14 (5): 513–20. https://doi.org/10.1038/nmeth.4256.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. "Genomic Insights That Advance the Species Definition for Prokaryotes." *Proceedings of the National Academy of Sciences* 102 (7): 2567–72. https://doi.org/10.1073/pnas.0409727102.

Koonin, E. V., and Y. I. Wolf. 2008. "Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World." *Nucleic Acids Research* 36 (21): 6688–719. https://doi.org/10.1093/nar/gkn668.

Koonin, Eugene V. 2005. "Orthologs, Paralogs, and Evolutionary Genomics." *Annual Review of Genetics* 39 (1): 309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725.

Koonin, Eugene V. 2011. "Are There Laws of Genome Evolution?" *PLoS Computational Biology* 7 (8): e1002173. https://doi.org/10.1371/journal.pcbi.1002173.

Koonin, Eugene V. 2015. "Origin of Eukaryotes from Within Archaea, Archaeal Eukaryome and Bursts of Gene Gain: Eukaryogenesis Just Made Easier?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1678): 20140333. https://doi.org/10.1098/rstb.2014.0333.

Koonin, Eugene V. 2016. "Horizontal Gene Transfer: Essentiality and Evolvability in Prokaryotes, and Roles in Evolutionary Transitions." *F1000Research* 5: 1805. https://doi.org/10.12688/f1000research.8737.1.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, et al. 2020. "Global Organization and Proposed Megataxonomy of the Virus World." *Microbiology and Molecular Biology Reviews* 84 (2). https://doi.org/10.1128/mmbr.00061-19.

Koonin, Eugene V., Kira S. Makarova, and L. Aravind. 2001. "Horizontal Gene Transfer in Prokaryotes: Quantification and Classification." *Annual Review of Microbiology* 55 (1): 709–42. https://doi.org/10.1146/annurev.micro.55.1.709.

Koonin, Eugene V., Yuri I. Wolf, and Georgy P. Karev. 2002. "The Structure of the Protein Universe and Genome Evolution." *Nature* 420 (6912): 218–23. https://doi.org/10.1038/nature01256.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive k -Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Korf, Ian. 2004. "Gene Finding in Novel Genomes." *BMC Bioinformatics* 5 (1). https://doi.org/10.1186/1471-2105-5-59.

Korsunsky, Ilya, Nghia Millard, Jean Fan, et al. 2019. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony." *Nature Methods* 16 (12): 1289–96. https://doi.org/10.1038/s41592-019-0619-0.

Kortemme, Tanja. 2024. "De Novo Protein Design—from New Structures to Programmable Functions." *Cell* 187 (3): 526–44. https://doi.org/10.1016/j.cell.2023.12.028.

Korthauer, Keegan, Patrick K. Kimes, Claire Duvallet, et al. 2019. "A Practical Guide to Methods Controlling False Discoveries in Computational Biology." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1716-1.

Kosakovsky Pond, Sergei L., David Posada, Michael B. Gravenor, Christopher H. Woelk, and Simon D. W. Frost. 2006. "Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm." *Molecular Biology and Evolution* 23 (10): 1891–901. https://doi.org/10.1093/molbev/msl051.

Kosiol, Carolin, and Maria Anisimova. 2019. "Selection Acting on Genomes." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_12.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake—a Scalable Bioinformatics Workflow Engine."

Bioinformatics 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Koutrouli, Mikaela, Evangelos Karatzas, David Paez-Espino, and Georgios A. Pavlopoulos. 2020. "A Guide to Conquer the Biological Network Era Using Graph Theory." *Frontiers in Bioengineering and Biotechnology* 8. https://doi.org/10.3389/fbioe.2020.00034.

Kovalevskiy, Oleg, Juan Mateos-Garcia, and Kathryn Tunyasuvunakool. 2024. *AlphaFold Two Years on: Validation and Impact.* https://doi.org/10.48550/arXiv.2403.02124.

Kozakov, Dima, David R Hall, Bing Xia, et al. 2017. "The ClusPro Web Server for Protein–Protein Docking." *Nature Protocols* 12 (2): 255–78. https://doi.org/10.1038/nprot.2016.169.

Krings, Matthias, Anne Stone, Ralf W Schmitz, Heike Krainitzki, Mark Stoneking, and Svante Pääbo. 1997. "Neandertal DNA Sequences and the Origin of Modern Humans." *Cell* 90 (1): 19–30. https://doi.org/10.1016/s0092-8674(00)80310-4.

Krishna, Rohith, Jue Wang, Woody Ahern, et al. 2024. "Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom." *Science* 384 (6693). https://doi.org/10.1126/science.adl2528.

Krivov, Georgii G., Maxim V. Shapovalov, and Roland L. Dunbrack. 2009. "Improved Prediction of Protein Side-Chain Conformations with SCWRL4." *Proteins: Structure, Function, and Bioinformatics* 77 (4): 778–95. https://doi.org/10.1002/prot.22488.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

Krogan, Nevan J., Gerard Cagney, Haiyuan Yu, et al. 2006. "Global Landscape of Protein Complexes in the Yeast Saccharomyces Cerevisiae." *Nature* 440 (7084): 637–43. https://doi.org/10.1038/nature04670.

Krogh, Anders. 2008. "What Are Artificial Neural Networks?" *Nature Biotechnology* 26 (2): 195–97. https://doi.org/10.1038/nbt1386.

Krogh, Anders, Michael Brown, I.Saira Mian, Kimmen Sjölander, and David Haussler. 1994. "Hidden Markov Models in Computational Biology." *Journal of Molecular Biology* 235 (5): 1501–31. https://doi.org/10.1006/jmbi.1994.1104.

Krogh, Anders, Björn Larsson, Gunnar von Heijne, and Erik L. L Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305 (3): 567–80. https://doi.org/10.1006/jmbi.2000.4315.

Krueger, Felix, and Simon R. Andrews. 2011. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." *Bioinformatics* 27 (11): 1571–72. https://doi.org/10.1093/bioinformatics/btr167.

Kruger, Kelly, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. 1982. "Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena." *Cell* 31 (1): 147–57. https://doi.org/10.1016/0092-8674(82)90414-7.

Kryazhimskiy, Sergey, and Joshua B. Plotkin. 2008. "The Population Genetics of dN/dS." *PLoS Genetics* 4 (12): e1000304. https://doi.org/10.1371/journal.pgen.1000304.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2021. "Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIV." *Proteins: Structure, Function, and Bioinformatics* 89 (12): 1607–17. https://doi.org/10.1002/prot.26237.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2023. "Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XV." *Proteins: Structure, Function, and Bioinformatics* 91 (12): 1539–49. https://doi.org/10.1002/prot.26617.

Krzywinski, Martin, and Naomi Altman. 2013. "Significance, p Values and t-Tests." *Nature Methods* 10 (11): 1041–42. https://doi.org/10.1038/nmeth.2698.

Kühlbrandt, Werner. 2014. "The Resolution Revolution." *Science* 343 (6178): 1443–44. https://doi.org/10.1126/science.1251652.

Kuhlman, Brian, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. 2003. "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy." *Science* 302 (5649): 1364–68. https://doi.org/10.1126/science.1089427.

Kuhn, Kenneth, Shawn C. Baker, Eugene Chudin, et al. 2004. "A Novel, High-Performance Random Array Platform for Quantitative Gene Expression Profiling." *Genome Research* 14 (11): 2347–56. https://doi.org/10.1101/gr.2739104.

Kulkarni, Prakash, Lauren Porter, Tsui-Fen Chou, et al. 2025. "Evolving Concepts of the Protein Universe." *iScience* 28 (3): 112012. https://doi.org/10.1016/j.isci.2025.112012.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22 (1): 79–86. https://doi.org/10.1214/aoms/1177729694.

Kulmanov, Maxat, and Robert Hoehndorf. 2019. "DeepGOPlus: Improved Protein Function Prediction from Sequence." *Bioinformatics* 36 (2): 422–29. https://doi.org/10.1093/bioinformatics/btz595.

Kumar, Manjeet, Sushama Michael, Jesús Alvarado-Valverde, et al. 2021. "The Eukaryotic Linear Motif Resource: 2022 Release." *Nucleic Acids Research* 50 (D1): D497–508. https://doi.org/10.1093/nar/gkab975.

Kummerfeld, Sarah K., and Sarah A. Teichmann. 2005. "Relative Rates of Gene Fusion and Fission in Multi-Domain Proteins." *Trends in Genetics* 21 (1): 25–30. https://doi.org/10.1016/j.tig.2004.11.007.

Kuntz, Irwin D., Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. 1982. "A Geometric Approach to Macromolecule-Ligand Interactions." *Journal of Molecular Biology* 161 (2): 269–88. https://doi.org/10.1016/0022-2836(82)90153-x.

Kurtz, Stefan, Adam Phillippy, Arthur L Delcher, et al. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome Biology* 5 (2). https://doi.org/10.1186/gb-2004-5-2-r12.

Kuznetsov, Dmitry, Fredrik Tegenfeldt, Mosè Manni, et al. 2022. "OrthoDB V11: Annotation of Orthologs in the Widest Sampling of Organismal Diversity." *Nucleic Acids Research* 51 (D1): D445–51. https://doi.org/10.1093/nar/gkac998.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, et al. 2018. "RNA Velocity of Single Cells." *Nature* 560 (7719): 494–98. https://doi.org/10.1038/s41586-018-0414-6.

Laine, Romain F., Ignacio Arganda-Carreras, Ricardo Henriques, and Guillaume Jacquemet. 2021. "Avoiding a Replication Crisis in Deep-Learning-Based Bioimage Analysis." *Nature Methods* 18 (10): 1136–44. https://doi.org/10.1038/s41592-021-01284-3.

Lam, Henry, Eric W. Deutsch, James S. Eddes, et al. 2007. "Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS." *PROTEOMICS* 7 (5): 655–67. https://doi.org/10.1002/pmic.200600625.

Lambert, Christophe, Jean-Marc Van Campenhout, Xavier DeBolle, and Eric Depiereux. 2003. "Review of Common Sequence Alignment Methods: Clues to Enhance Reliability." *Current Genomics* 4: 131–46.

Lamm, Lorenz, Simon Zufferey, Hanyi Zhang, et al. 2026. "MemBrain V2: An End-to-End Tool for the Analysis of Membranes in Cryo-Electron Tomography." *Nature Methods*, ahead of print. https://doi.org/10.1038/s41592-026-03178-8.

Lander, Eric S., and David Botstein. 1987. "Homozygosity Mapping: A Way to Map Human Recessive Traits with the DNA of Inbred Children." *Science* 236 (4808): 1567–70. https://doi.org/10.1126/science.2884728.

Lander, Eric S., and Michael S. Waterman. 1988. "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis." *Genomics* 2 (3): 231–39. https://doi.org/10.1016/0888-7543(88)90007-9.

Lander, Eric, and Leonid Kruglyak. 1995. "Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results." *Nature Genetics* 11 (3): 241–47. https://doi.org/10.1038/ng1195-241.

Landrum, Melissa J, Jennifer M Lee, Mark Benson, et al. 2017. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research* 46 (D1): D1062–67. https://doi.org/10.1093/nar/gkx1153.

Landry, Christian R., Emmanuel D. Levy, and Stephen W. Michnick. 2009. "Weak Functional Constraints on Phosphoproteomes." *Trends in Genetics* 25 (5): 193–97. https://doi.org/10.1016/j.tig.2009.03.003.

Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." *Genome Research* 22 (9): 1813–31. https://doi.org/10.1101/gr.136184.111.

Lang, B. Franz, Gertraud Burger, Charles J. O'Kelly, et al. 1997. "An Ancestral Mitochondrial DNA Resembling a Eubacterial Genome in Miniature." *Nature* 387 (6632): 493–97. https://doi.org/10.1038/387493a0.

Lange, Vinzenz, Paola Picotti, Bruno Domon, and Ruedi Aebersold. 2008. "Selected Reaction Monitoring for Quantitative Proteomics: A Tutorial." *Molecular Systems Biology* 4 (1). https://doi.org/10.1038/msb.2008.61.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An r Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1). https://doi.org/10.1186/1471-2105-9-559.

Langley, Charles H., and Walter M. Fitch. 1974. "An Examination of the Constancy of the Rate of Molecular Evolution." *Journal of Molecular Evolution* 3: 161–77.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3). https://doi.org/10.1186/gb-2009-10-3-r25.

Lapidus, Alla L., and Anton I. Korobeynikov. 2021. "Metagenomic Data Assembly – the Way of Decoding Unknown Microorganisms." *Frontiers in Microbiology* 12. https://doi.org/10.3389/fmicb.2021.613791.

Lartillot, Nicolas, and Hervé Philippe. 2004. "A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process." *Molecular Biology and Evolution* 21 (6): 1095–109. https://doi.org/10.1093/molbev/msh112.

Laskowski, R. A., J. D. Watson, and J. M. Thornton. 2005. "ProFunc: A Server for Predicting Protein Function from 3D Structure." *Nucleic Acids Research* 33 (Web Server): W89–93. https://doi.org/10.1093/nar/gki414.

Lau, Andy M., Nicola Bordin, Shaun M. Kandathil, et al. 2024. "Exploring Structural Diversity Across the Protein Universe with the Encyclopedia of Domains." *Science* 386 (6721). https://doi.org/10.1126/science.adq4946.

Laurents, Douglas V. 2022. "AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function." *Frontiers in Molecular Biosciences* 9. https://doi.org/10.3389/fmolb.2022.906437.

Law, Charity W., Monther Alhamdoosh, Shian Su, et al. 2018. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and edgeR." *F1000Research* 5: 1408. https://doi.org/10.12688/f1000research.9005.3.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2). https://doi.org/10.1186/gb-2014-15-2-r29.

Lawrence, Charles E., Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton. 1993. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment." *Science* 262 (5131): 208–14. https://doi.org/10.1126/science.8211139.

Lawrence, Michael S., Petar Stojanov, Paz Polak, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457): 214–18. https://doi.org/10.1038/nature12213.

Lawson, Devon A., Kai Kessenbrock, Ryan T. Davis, Nicholas Pervolarakis, and Zena Werb. 2018. "Tumour Heterogeneity and Metastasis at Single-Cell Resolution." *Nature Cell Biology* 20 (12): 1349–60. https://doi.org/10.1038/s41556-018-0236-7.

Lazar, Cosmin, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. 2016. "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies." *Journal of Proteome Research* 15 (4): 1116–25. https://doi.org/10.1021/acs.jproteome.5b00981.

Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery. 2009. "Fpocket: An Open Source Platform for Ligand Pocket Detection." *BMC Bioinformatics* 10 (1). https://doi.org/10.1186/1471-2105-10-168.

Le, S. Q., and O. Gascuel. 2008. "An Improved General Amino Acid Replacement Matrix." *Molecular*

Biology and Evolution 25 (7): 1307–20. https://doi.org/10.1093/molbev/msn067.

Le, Trang, Casper F. Winsnes, Ulrika Axelsson, et al. 2022. "Analysis of the Human Protein Atlas Weakly Supervised Single-Cell Classification Competition." *Nature Methods* 19 (10): 1221–29. https://doi.org/10.1038/s41592-022-01606-z.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–324. https://doi.org/10.1109/5.726791.

Lee, Benjamin D., Anthony Gitter, Casey S. Greene, et al. 2022. "Ten Quick Tips for Deep Learning in Biology." *PLOS Computational Biology* 18 (3): e1009803. https://doi.org/10.1371/journal.pcbi.1009803.

Lee, Insuk, Shailesh V. Date, Alex T. Adai, and Edward M. Marcotte. 2004. "A Probabilistic Functional Network of Yeast Genes." *Science* 306 (5701): 1555–58. https://doi.org/10.1126/science.1099511.

Lee, Insuk, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G Fraser, and Edward M Marcotte. 2008. "A Single Gene Network Accurately Predicts Phenotypic Effects of Gene Perturbation in Caenorhabditis Elegans." *Nature Genetics* 40 (2): 181–88. https://doi.org/10.1038/ng.2007.70.

Lee, Jennifer A., Claudia M. B. Carvalho, and James R. Lupski. 2007. "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders." *Cell* 131 (7): 1235–47. https://doi.org/10.1016/j.cell.2007.11.037.

Lee, Robin van der, Marija Buljan, Benjamin Lang, et al. 2014. "Classification of Intrinsically Disordered Regions and Proteins." *Chemical Reviews* 114 (13): 6589–631. https://doi.org/10.1021/cr400525m.

Lee, Rosalind C., Rhonda L. Feinbaum, and Victor Ambros. 1993. "The c. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14." *Cell* 75 (5): 843–54. https://doi.org/10.1016/0092-8674(93)90529-y.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, et al. 2010. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." *Nature Reviews Genetics* 11 (10): 733–39. https://doi.org/10.1038/nrg2825.

Legrain, Pierre, Ruedi Aebersold, Alexander Archakov, et al. 2011. "The Human Proteome Project: Current State and Future Direction." *Molecular & Cellular Proteomics* 10 (7): M111.009993. https://doi.org/10.1074/mcp.m111.009993.

Leman, Julia Koehler, Brian D. Weitzner, Steven M. Lewis, et al. 2020. "Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks." *Nature Methods* 17 (7): 665–80. https://doi.org/10.1038/s41592-020-0848-2.

Lemoine, F., J.-B. Domelevo Entfellner, E. Wilkinson, et al. 2018. "Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data." *Nature* 556 (7702): 452–56. https://doi.org/10.1038/s41586-018-0043-0.

Lever, Jake, Martin Krzywinski, and Naomi Altman. 2016. "Model Selection and Overfitting." *Nature Methods* 13 (9): 703–4. https://doi.org/10.1038/nmeth.3968.

Lever, Jake, Martin Krzywinski, and Naomi Altman. 2017. "Principal Component Analysis." *Nature Methods* 14 (7): 641–42. https://doi.org/10.1038/nmeth.4346.

Levinthal, Cyrus. 1966. "Molecular Model-Building by Computer." *Scientific American* 214 (6): 42–52. https://doi.org/10.1038/scientificamerican0666-42.

Levitt, Michael, and Cyrus Chothia. 1976. "Structural Patterns in Globular Proteins." *Nature* 261 (5561): 552–58. https://doi.org/10.1038/261552a0.

Levitt, Michael, and Arieh Warshel. 1975. "Computer Simulation of Protein Folding." *Nature* 253 (5494): 694–98. https://doi.org/10.1038/253694a0.

Levy, Emmanuel D., Elisabetta Boeri Erba, Carol V. Robinson, and Sarah A. Teichmann. 2008. "Assembly Reflects Evolution of Protein Complexes." *Nature* 453 (7199): 1262–65. https://doi.org/10.1038/nature06942.

Li, Bo, and Colin N Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or Without a Reference Genome." *BMC Bioinformatics* 12 (1). https://doi.org/10.1186/1471-2105-12-323.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

Li, H. 2006. "TreeFam: A Curated Database of Phylogenetic Trees of Animal Gene Families." *Nucleic Acids Research* 34 (90001): D572–80. https://doi.org/10.1093/nar/gkj118.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, and Richard Durbin. 2010. "Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 26 (5): 589–95. https://doi.org/10.1093/bioinformatics/btp698.

Li, Heng, Bob Handsaker, Alec Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. "Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores." *Genome Research* 18 (11): 1851–58. https://doi.org/10.1101/gr.078212.108.

Li, Li, Christian J. Stoeckert, and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.

Li, Qing, Zhihang Hu, Yixuan Wang, et al. 2024. "Progress and Opportunities of Foundation Models in Bioinformatics." *Briefings in Bioinformatics* 25 (6). https://doi.org/10.1093/bib/bbae548.

Li, Qunhua, James B. Brown, Haiyan Huang, and Peter J. Bickel. 2011. "Measuring Reproducibility of High-Throughput Experiments." *The Annals of Applied Statistics* 5 (3). https://doi.org/10.1214/11-aoas466.

Li, Wei, Han Xu, Tengfei Xiao, et al. 2014. "MAGeCK Enables Robust Identification of Essential Genes from Genome-Scale CRISPR/Cas9 Knockout Screens." *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0554-4.

Li, Weizhong, and Adam Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Liao, Wen-Wei, Mobin Asri, Jana Ebler, et al. 2023. "A Draft Human Pangenome Reference." *Nature* 617 (7960): 312–24. https://doi.org/10.1038/s41586-023-05896-x.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database Hallmark Gene Set Collection." *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.

Licatalosi, Donny D., Aldo Mele, John J. Fak, et al. 2008. "HITS-CLIP Yields Genome-Wide Insights into Brain Alternative RNA Processing." *Nature* 456 (7221): 464–69. https://doi.org/10.1038/nature07488.

Lieber, Michael R. 2010. "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway." *Annual Review of Biochemistry* 79 (1): 181–211. https://doi.org/10.1146/annurev.biochem.052308.093131.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Light, S., R. Sagit, O. Sachenkova, D. Ekman, and A. Elofsson. 2013. "Protein Expansion Is Primarily Due to Indels in Intrinsically Disordered Regions." *Molecular Biology and Evolution* 30 (12): 2645–53. https://doi.org/10.1093/molbev/mst157.

Lima-Mendez, Gipsi, and Jacques van Helden. 2009. "The Powerful Law of the Power Law and Other Myths in Network Biology." *Molecular BioSystems* 5 (12): 1482–93. https://doi.org/10.1039/b908681a.

Lin, Baohui, Xiaoling Luo, Yumeng Liu, and Xiaopeng Jin. 2024. "A Comprehensive Review and Comparison of Existing Computational Methods for Protein Function Prediction." *Briefings in Bioinformatics* 25 (4). https://doi.org/10.1093/bib/bbae289.

Lin, Michael F., Irwin Jungreis, and Manolis Kellis. 2011. "PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-Coding Regions." *Bioinformatics* 27 (13): i275–82. https://doi.org/10.1093/bioinformatics/btr209.

Lin, Pei-Yu, Shiang-Chin Huang, Kuan-Lin Chen, et al. 2025. "Analysing Protein Complexes in Plant Science: Insights and Limitation with AlphaFold 3." *Botanical Studies* 66 (1). https://doi.org/10.1186/s40529-025-00462-2.

Lin, Zeming, Halil Akin, Roshan Rao, et al. 2023. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." *Science* 379 (6637): 1123–30. https://doi.org/10.1126/science.ade2

574.

Lincoln, Tracey A., and Gerald F. Joyce. 2009. "Self-Sustained Replication of an RNA Enzyme." *Science* 323 (5918): 1229–32. https://doi.org/10.1126/science.1167856.

Lindahl, Erik, and Arne Elofsson. 2000. "Identification of Related Proteins on Family, Superfamily and Fold Level." *Journal of Molecular Biology* 295 (3): 613–25. https://doi.org/10.1006/jmbi.1999.3377.

Lindorff-Larsen, Kresten, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. 2012. "Systematic Validation of Protein Force Fields Against Experimental Data." *PLoS ONE* 7 (2): e32131. https://doi.org/10.1371/journal.pone.0032131.

Lindorff-Larsen, Kresten, Stefano Piana, Ron O. Dror, and David E. Shaw. 2011. "How Fast-Folding Proteins Fold." *Science* 334 (6055): 517–20. https://doi.org/10.1126/science.1208351.

Lipman, D J, S F Altschul, and J D Kececioglu. 1989. "A Tool for Multiple Sequence Alignment." *Proceedings of the National Academy of Sciences* 86 (12): 4412–15. https://doi.org/10.1073/pnas.86.12.4412.

Lipman, David J., and William R. Pearson. 1985. "Rapid and Sensitive Protein Similarity Searches." *Science* 227 (4693): 1435–41. https://doi.org/10.1126/science.2983426.

Lister, Ryan, Mattia Pelizzola, Robert H. Dowen, et al. 2009. "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences." *Nature* 462 (7271): 315–22. https://doi.org/10.1038/nature08514.

Littmann, Maria, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost. 2021. "Embeddings from Deep Learning Transfer GO Annotations Beyond Homology." *Scientific Reports* 11 (1). https://doi.org/10.1038/s41598-020-80786-0.

Liu, Dang, Martin Hunt, and Isheng J Tsai. 2018. "Inferring Synteny Between Genome Assemblies: A Systematic Evaluation." *BMC Bioinformatics* 19 (1). https://doi.org/10.1186/s12859-018-2026-4.

Liu, Liang, Christian Anderson, Dennis Pearl, and Scott V. Edwards. 2019. "Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_7.

Liu, Wei, Ziye Wang, Ronghui You, et al. 2024. "PLMSearch: Protein Language Model Powers Accurate and Fast Sequence Search for Remote Homology." *Nature Communications* 15 (1). https://doi.org/10.1038/s41467-024-46808-5.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on mRNA Abundance." *Cell* 165 (3): 535–50. https://doi.org/10.1016/j.cell.2016.03.014.

Livesey, Benjamin J, and Joseph A Marsh. 2023. "Updated Benchmarking of Variant Effect Predictors Using Deep Mutational Scanning." *Molecular Systems Biology* 19 (8). https://doi.org/10.15252/msb.202211474.

Ljosa, Vebjorn, Katherine L Sokolnicki, and Anne E Carpenter. 2012. "Annotated High-Throughput Microscopy Image Sets for Validation." *Nature Methods* 9 (7): 637–37. https://doi.org/10.1038/nmeth.2083.

Lobry, J. R. 1996. "Asymmetric Substitution Patterns in the Two DNA Strands of Bacteria." *Molecular Biology and Evolution* 13 (5): 660–65. https://doi.org/10.1093/oxfordjournals.molbev.a025626.

Lockhart, David J., Helin Dong, Michael C. Byrne, et al. 1996. "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." *Nature Biotechnology* 14 (13): 1675–80. https://doi.org/10.1038/nbt1296-1675.

Loh, Po-Ru, George Tucker, Brendan K Bulik-Sullivan, et al. 2015. "Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts." *Nature Genetics* 47 (3): 284–90. https://doi.org/10.1038/ng.3190.

Long, Manyuan, Esther Betrán, Kevin Thornton, and Wen Wang. 2003. "The Origin of New Genes: Glimpses from the Young and Old." *Nature Reviews Genetics* 4 (11): 865–75. https://doi.org/10.1038/nrg1204.

Lopez, Romain, Adam Gayoso, and Nir Yosef. 2020. "Enhancing Scientific Discoveries in Molecular Biology with Deep Generative Models." *Molecular Systems Biology* 16 (9). https://doi.org/10.15252/msb.20199198.

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. "Deep Generative Modeling for Single-Cell Transcriptomics." *Nature Methods* 15 (12): 1053–58. https://doi.org/10.1038/s41592-018-0229-2.

Love, Michael I., Simon Anders, Vladislav Kim, and Wolfgang Huber. 2016. "RNA-Seq Workflow: Gene-Level Exploratory Analysis and Differential Expression." *F1000Research* 4: 1070. https://doi.org/10.12688/f1000research.7035.2.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

Lovén, Jakob, David A. Orlando, Alla A. Sigova, et al. 2012. "Revisiting Global Gene Expression Analysis." *Cell* 151 (3): 476–82. https://doi.org/10.1016/j.cell.2012.10.012.

Lowe, Rohan, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. 2017. "Transcriptomics Technologies." *PLOS Computational Biology* 13 (5): e1005457. https://doi.org/10.1371/journal.pcbi.1005457.

Lowe, Todd M., and Sean R. Eddy. 1997. "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." *Nucleic Acids Research* 25 (5): 955–64. https://doi.org/10.1093/nar/25.5.955.

Löytynoja, Ari, and Nick Goldman. 2008. "Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis." *Science* 320 (5883): 1632–35. https://doi.org/10.1126/science.1158395.

Lu, Ming Y., Bowen Chen, Drew F. K. Williamson, et al. 2024. "A Visual-Language Foundation Model for Computational Pathology." *Nature Medicine* 30 (3): 863–74. https://doi.org/10.1038/s41591-024-02856-4.

Lubiana, Tiago, Rafael Lopes, Pedro Medeiros, et al. 2023. "Ten Quick Tips for Harnessing the Power

of ChatGPT in Computational Biology." *PLOS Computational Biology* 19 (8): e1011319. https://doi.org/10.1371/journal.pcbi.1011319.

Luck, Katja, Dae-Kyum Kim, Luke Lambourne, et al. 2020. "A Reference Map of the Human Binary Protein Interactome." *Nature* 580 (7803): 402–8. https://doi.org/10.1038/s41586-020-2188-x.

Ludwig, Christina, Ludovic Gillet, George Rosenberger, Sabine Amon, Ben C Collins, and Ruedi Aebersold. 2018. "Data-Independent Acquisition-Based SWATH-MS for Quantitative Proteomics: A Tutorial." *Molecular Systems Biology* 14 (8). https://doi.org/10.15252/msb.20178126.

Luecken, Malte D., M. Büttner, K. Chaichoompu, et al. 2021. "Benchmarking Atlas-Level Data Integration in Single-Cell Genomics." *Nature Methods* 19 (1): 41–50. https://doi.org/10.1038/s41592-021-01336-8.

Luecken, Malte D, and Fabian J Theis. 2019. "Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial." *Molecular Systems Biology* 15 (6). https://doi.org/10.15252/msb.20188746.

Lukashin, Alexander V., and Mark Borodovsky. 1998. "GeneMark.hmm: New Solutions for Gene Finding." *Nucleic Acids Research* 26 (4): 1107–15. https://doi.org/10.1093/nar/26.4.1107.

Lun, Aaron T. L., Karsten Bach, and John C. Marioni. 2016. "Pooling Across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0947-7.

Lund, Steven P., Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. 2012. "Detecting Differential Expression in RNA-Sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates." *Statistical Applications in Genetics and Molecular Biology* 11 (5). https://doi.org/10.1515/1544-6115.1826.

Lundberg, Emma, and Georg H. H. Borner. 2019. "Spatial Proteomics: A Powerful Discovery Tool for Cell Biology." *Nature Reviews Molecular Cell Biology* 20 (5): 285–302. https://doi.org/10.1038/s41580-018-0094-y.

Lupski, James R. 1998. "Genomic Disorders: Structural Features of the Genome Can Lead to DNA Rearrangements and Human Disease Traits." *Trends in Genetics* 14 (10): 417–22. https://doi.org/10.1016/s0168-9525(98)01555-8.

Lussi, Yvonne C., Michele Magrane, Maria J. Martin, and Sandra and Orchard. 2023. "Searching and Navigating UniProt Databases." *Current Protocols* 3 (3). https://doi.org/10.1002/cpz1.700.

Lynch, Michael, Matthew S. Ackerman, Jean-Francois Gout, et al. 2016. "Genetic Drift, Selection and the Evolution of the Mutation Rate." *Nature Reviews Genetics* 17 (11): 704–14. https://doi.org/10.1038/nrg.2016.104.

Lynch, Michael, and John S. Conery. 2000. "The Evolutionary Fate and Consequences of Duplicate Genes." *Science* 290 (5494): 1151–55. https://doi.org/10.1126/science.290.5494.1151.

Lynch, Michael, and John S. Conery. 2003. "The Origins of Genome Complexity." *Science* 302 (5649): 1401–4. https://doi.org/10.1126/science.1089370.

Ma, Bin, Kaizhong Zhang, Christopher Hendrie, et al. 2003. "PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry." *Rapid Communications in Mass Spectrometry* 17

(20): 2337–42. https://doi.org/10.1002/rcm.1196.

Ma, Shaoqian, and Yongyou Zhang. 2020. "Profiling Chromatin Regulatory Landscape: Insights into the Development of ChIP-Seq and ATAC-Seq." *Molecular Biomedicine* 1 (1). https://doi.org/10.1186/s43556-020-00009-w.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–605. https://www.jmlr.org/papers/v9/vandermaaten08a.html.

Macosko, Evan Z., Anindita Basu, Rahul Satija, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Madani, Ali, Ben Krause, Eric R. Greene, et al. 2023. "Large Language Models Generate Functional Protein Sequences Across Diverse Families." *Nature Biotechnology* 41 (8): 1099–106. https://doi.org/10.1038/s41587-022-01618-2.

Maddison, Wayne P. 1997. "Gene Trees in Species Trees." *Systematic Biology* 46 (3): 523–36. https://doi.org/10.1093/sysbio/46.3.523.

Madeira, S. C., and A. L. Oliveira. 2004. "Biclustering Algorithms for Biological Data Analysis: A Survey." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1 (1): 24–45. https://doi.org/10.1109/tcbb.2004.2.

Mahdessian, Diana, Anthony J. Cesnik, Christian Gnann, et al. 2021. "Spatiotemporal Dissection of the Cell Cycle with Single-Cell Proteogenomics." *Nature* 590 (7847): 649–54. https://doi.org/10.1038/s41586-021-03232-9.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1828-7.

Makałowski, Wojciech, Valer Gotea, Amit Pande, and Izabela Makałowska. 2019. "Transposable Elements: Classification, Identification, and Their Use as a Tool for Comparative Genomics." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_6.

Makarov, Alexander. 2000. "Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis." *Analytical Chemistry* 72 (6): 1156–62. https://doi.org/10.1021/ac991131p.

Malik-Sheriff, Rahuman S, Mihai Glont, Tung V N Nguyen, et al. 2019. "BioModels—15 Years of Sharing Computational Models in Life Science." *Nucleic Acids Research*, ahead of print. https://doi.org/10.1093/nar/gkz1055.

Manber, Udi, and Gene Myers. 1993. "Suffix Arrays: A New Method for on-Line String Searches." *SIAM Journal on Computing* 22 (5): 935–48. https://doi.org/10.1137/0222058.

Mandell, Daniel J, Evangelos A Coutsias, and Tanja Kortemme. 2009. "Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling." *Nature Methods* 6 (8): 551–52. https://doi.org/10.1038/nmeth0809-551.

Mangan, S., and U. Alon. 2003. "Structure and Function of the Feed-Forward Loop Network Motif." *Proceedings of the National Academy of Sciences* 100 (21): 11980–85. https://doi.org/10.1073/pnas.2133841100.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53. https://doi.org/10.1038/nature08494.

MAQC Consortium. 2006. "The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements." *Nature Biotechnology* 24 (9): 1151–61. https://doi.org/10.1038/nbt1239.

Marbach, Daniel, James C. Costello, Robert Küffner, et al. 2012. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9 (8): 796–804. https://doi.org/10.1038/nmeth.2016.

Marchese, Francesco P., Ivan Raimondi, and Maite Huarte. 2017. "The Multidimensional Mechanisms of Long Noncoding RNA Function." *Genome Biology* 18 (1). https://doi.org/10.1186/s13059-017-1348-2.

Marchini, Jonathan, and Bryan Howie. 2010. "Genotype Imputation for Genome-Wide Association Studies." *Nature Reviews Genetics* 11 (7): 499–511. https://doi.org/10.1038/nrg2796.

Marcotte, Edward M., Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, and David Eisenberg. 1999. "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences." *Science* 285 (5428): 751–53. https://doi.org/10.1126/science.285.5428.751.

Marcotte, Edward M., Matteo Pellegrini, Michael J. Thompson, Todd O. Yeates, and David Eisenberg. 1999. "A Combined Algorithm for Genome-Wide Prediction of Protein Function." *Nature* 402 (6757): 83–86. https://doi.org/10.1038/47048.

Mardikoraem, Mehrsa, Zirui Wang, Nathaniel Pascual, and Daniel Woldring. 2023. "Generative Models for Protein Sequence Modeling: Recent Advances and Future Directions." *Briefings in Bioinformatics* 24 (6). https://doi.org/10.1093/bib/bbad358.

Margolin, Adam A, Ilya Nemenman, Katia Basso, et al. 2006. "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context." *BMC Bioinformatics* 7 (S1). https://doi.org/10.1186/1471-2105-7-s1-s7.

Margulies, Marcel, Michael Egholm, William E. Altman, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80. https://doi.org/10.1038/nature03959.

Margush, T., and F. R. McMorris. 1981. "Consensusn-Trees." *Bulletin of Mathematical Biology* 43 (2): 239–44. https://doi.org/10.1007/bf02459446.

Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. "lDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests." *Bioinformatics* 29 (21): 2722–28. https://doi.org/10.1093/bioinformatics/btt473.

Marioni, John C., Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18 (9): 1509–17. https://doi.org/10.1101/gr.079558.108.

Maritan, Martina, Ludovic Autin, Jonathan Karr, Markus W. Covert, Arthur J. Olson, and David S. Goodsell. 2022. "Building Structural Models of a Whole Mycoplasma Cell." *Journal of Molecular Biology* 434 (2): 167351. https://doi.org/10.1016/j.jmb.2021.167351.

Marks, Debora S., Lucy J. Colwell, Robert Sheridan, et al. 2011. "Protein 3D Structure Computed from Evolutionary Sequence Variation." *PLoS ONE* 6 (12): e28766. https://doi.org/10.1371/journal.pone.0028766.

Marks, Debora S, Thomas A Hopf, and Chris Sander. 2012. "Protein Structure Prediction from Sequence Variation." *Nature Biotechnology* 30 (11): 1072–80. https://doi.org/10.1038/nbt.2419.

Marks, Markus, Uriah Israel, Rohit Dilip, et al. 2025. "CellSAM: A Foundation Model for Cell Segmentation." *Nature Methods* 22 (12): 2585–93. https://doi.org/10.1038/s41592-025-02879-w.

Marku, Malvina, and Vera Pancaldi. 2023. "From Time-Series Transcriptomics to Gene Regulatory Networks: A Review on Inference Methods." *PLOS Computational Biology* 19 (8): e1011254. https://doi.org/10.1371/journal.pcbi.1011254.

Martens, Lennart, Matthew Chambers, Marc Sturm, et al. 2011. "mzML—a Community Standard for Mass Spectrometry Data." *Molecular & Cellular Proteomics* 10 (1): R110.000133. https://doi.org/10.1074/mcp.r110.000133.

Marth, Gabor T., Ian Korf, Mark D. Yandell, et al. 1999. "A General Approach to Single-Nucleotide Polymorphism Discovery." *Nature Genetics* 23: 452–56.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Martin, William, and Michael J. Russell. 2003. "On the Origins of Cells: A Hypothesis for the Evolutionary Transitions from Abiotic Geochemistry to Chemoautotrophic Prokaryotes, and from Prokaryotes to Nucleated Cells." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358 (1429): 59–85. https://doi.org/10.1098/rstb.2002.1183.

Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, et al. 2017. "Universal Patterns of Selection in Cancer and Somatic Tissues." *Cell* 171 (5): 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.

Marx, Vivien. 2021. "Method of the Year: Spatially Resolved Transcriptomics." *Nature Methods* 18 (1): 9–14. https://doi.org/10.1038/s41592-020-01033-y.

Maška, Martin, Vladimír Ulman, Pablo Delgado-Rodriguez, et al. 2023. "The Cell Tracking Challenge: 10 Years of Objective Benchmarking." *Nature Methods* 20 (7): 1010–20. https://doi.org/10.1038/s41592-023-01879-y.

Maslov, Sergei, and Kim Sneppen. 2002. "Specificity and Stability in Topology of Protein Networks." *Science* 296 (5569): 910–13. https://doi.org/10.1126/science.1065103.

Masson, Glenn R., John E. Burke, Natalie G. Ahn, et al. 2019. "Recommendations for Performing, Interpreting and Reporting Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) Experiments." *Nature Methods* 16 (7): 595–602. https://doi.org/10.1038/s41592-019-0459-y.

Mattick, John S., Paulo P. Amaral, Piero Carninci, et al. 2023. "Long Non-Coding RNAs: Definitions, Functions, Challenges and Recommendations." *Nature Reviews Molecular Cell Biology* 24 (6): 430–47. https://doi.org/10.1038/s41580-022-00566-8.

Maurer-Stroh, Sebastian, Maja Debulpaep, Nico Kuemmerer, et al. 2010. "Exploring the Sequence Determinants of Amyloid Structure Using Position-Specific Scoring Matrices." *Nature Methods* 7 (3): 237–42. https://doi.org/10.1038/nmeth.1432.

Maxam, A M, and W Gilbert. 1977. "A New Method for Sequencing DNA." *Proceedings of the National Academy of Sciences* 74 (2): 560–64. https://doi.org/10.1073/pnas.74.2.560.

Mayor, Michel, and Didier Queloz. 1995. "A Jupiter-Mass Companion to a Solar-Type Star." *Nature* 378 (6555): 355–59. https://doi.org/10.1038/378355a0.

McCammon, J. Andrew, Bruce R. Gelin, and Martin Karplus. 1977. "Dynamics of Folded Proteins." *Nature* 267 (5612): 585–90. https://doi.org/10.1038/267585a0.

McCarthy, Davis J., and Gordon K. Smyth. 2009. "Testing Significance Relative to a Fold-Change Threshold Is a TREAT." *Bioinformatics* 25 (6): 765–71. https://doi.org/10.1093/bioinformatics/btp053.

McClintock, Barbara. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences* 36 (6): 344–55. https://doi.org/10.1073/pnas.36.6.344.

McDonald, John H., and Martin Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

McGranahan, Nicholas, and Charles Swanton. 2017. "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future." *Cell* 168 (4): 613–28. https://doi.org/10.1016/j.cell.2017.01.018.

McHugh, Leo, and Jonathan W Arthur. 2008. "Computational Methods for Protein Identification from Mass Spectrometry Data." *PLoS Computational Biology* 4 (2): e12. https://doi.org/10.1371/journal.pcbi.0040012.

McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. "Why Prokaryotes Have Pangenomes." *Nature Microbiology* 2 (4). https://doi.org/10.1038/nmicrobiol.2017.40.

McInnes, Leland, John Healy, and James Melville. 2018. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://arxiv.org/abs/1802.03426.

McKenna, Aaron, Matthew Hanna, Eric Banks, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–303. https://doi.org/10.1101/gr.107524.110.

McLaren, William, Laurent Gil, Sarah E. Hunt, et al. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0974-4.

Medini, Duccio, Claudio Donati, Rino Rappuoli, and Hervé Tettelin. 2020. "The Pangenome: A Data-Driven Discovery in Biology." In *The Pangenome*. Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_1.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. "The

Microbial Pan-Genome." *Current Opinion in Genetics & Development* 15 (6): 589–94. https://doi.org/10.1016/j.gde.2005.09.006.

Meier, Florian, Andreas-David Brunner, Max Frank, et al. 2020. "diaPASEF: Parallel Accumulation–Serial Fragmentation Combined with Data-Independent Acquisition." *Nature Methods* 17 (12): 1229–36. https://doi.org/10.1038/s41592-020-00998-0.

Meier, Florian, Andreas-David Brunner, Scarlet Koch, et al. 2018. "Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer." *Molecular & Cellular Proteomics* 17 (12): 2534–45. https://doi.org/10.1074/mcp.tir118.000900.

Meier, Joshua, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. 2021. "Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2021.07.09.450648.

Meldal, Birgit H M, Livia Perfetto, Colin Combe, et al. 2021. "Complex Portal 2022: New Curation Frontiers." *Nucleic Acids Research* 50 (D1): D578–86. https://doi.org/10.1093/nar/gkab991.

Mellacheruvu, Dattatreya, Zachary Wright, Amber L Couzens, et al. 2013. "The CRAPome: A Contaminant Repository for Affinity Purification–Mass Spectrometry Data." *Nature Methods* 10 (8): 730–36. https://doi.org/10.1038/nmeth.2557.

Menche, Jörg, Amitabh Sharma, Maksim Kitsak, et al. 2015. "Uncovering Disease-Disease Relationships Through the Incomplete Interactome." *Science* 347 (6224). https://doi.org/10.1126/science.1257601.

Mendes, Fábio K, Dan Vanderpool, Ben Fulton, and Matthew W Hahn. 2020. "CAFE 5 Models Variation in Evolutionary Rates Among Gene Families." *Bioinformatics* 36 (22-23): 5516–18. https://doi.org/10.1093/bioinformatics/btaa1022.

Meng, Elaine C., Brian K. Shoichet, and Irwin D. Kuntz. 1992. "Automated Docking with Grid-Based Energy Evaluation." *Journal of Computational Chemistry* 13 (4): 505–24. https://doi.org/10.1002/jcc.540130412.

Mering, Christian von, Roland Krause, Berend Snel, et al. 2002. "Comparative Assessment of Large-Scale Data Sets of Protein–Protein Interactions." *Nature* 417 (6887): 399–403. https://doi.org/10.1038/nature750.

Merkin, Jason, Caitlin Russell, Ping Chen, and Christopher B. Burge. 2012. "Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues." *Science* 338 (6114): 1593–99. https://doi.org/10.1126/science.1228186.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92. https://doi.org/10.1063/1.1699114.

Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, et al. 2022. "Critical Assessment of Metagenome Interpretation: The Second Round of Challenges." *Nature Methods* 19 (4): 429–40. https://doi.org/10.1038/s41592-022-01431-4.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, et al. 2012. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual." *Science* 338 (6104): 222–26. https://doi.org/10.112

6/science.1224344.

Meyers, Robin M, Jordan G Bryan, James M McFarland, et al. 2017. "Computational Correction of Copy Number Effect Improves Specificity of CRISPR–Cas9 Essentiality Screens in Cancer Cells." *Nature Genetics* 49 (12): 1779–84. https://doi.org/10.1038/ng.3984.

Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. 2018. "PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools." *Nucleic Acids Research* 47 (D1): D419–26. https://doi.org/10.1093/nar/gky1038.

Milacic, Marija, Deidre Beavers, Patrick Conley, et al. 2023. "The Reactome Pathway Knowledgebase 2024." *Nucleic Acids Research* 52 (D1): D672–78. https://doi.org/10.1093/nar/gkad1025.

Miller, Catriona, Theo Portlock, Denis M. Nyaga, and Justin M. O'Sullivan. 2024. "A Review of Model Evaluation Metrics for Machine Learning in Genetics and Genomics." *Frontiers in Bioinformatics* 4. https://doi.org/10.3389/fbinf.2024.1457619.

Miller, Stanley L. 1953. "A Production of Amino Acids Under Possible Primitive Earth Conditions." *Science* 117 (3046): 528–29. https://doi.org/10.1126/science.117.3046.528.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (5594): 824–27. https://doi.org/10.1126/science.298.5594.824.

Minh, Bui Quang, Matthew W Hahn, and Robert Lanfear. 2020. "New Methods to Calculate Concordance Factors for Phylogenomic Datasets." *Molecular Biology and Evolution* 37 (9): 2727–33. https://doi.org/10.1093/molbev/msaa106.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, et al. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Minsky, Marvin, and Seymour Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Minussi, Darlan C., Michael D. Nicholson, Hanghui Ye, et al. 2021. "Breast Tumours Maintain a Reservoir of Subclonal Diversity During Expansion." *Nature* 592 (7853): 302–8. https://doi.org/10.1038/s41586-021-03357-x.

Mirarab, S., R. Reaz, Md. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics* 30 (17): i541–48. https://doi.org/10.1093/bioinformatics/btu462.

Mirdita, Milot, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes Söding, and Martin Steinegger. 2016. "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments." *Nucleic Acids Research* 45 (D1): D170–76. https://doi.org/10.1093/nar/gkw1081.

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. "ColabFold: Making Protein Folding Accessible to All." *Nature Methods* 19 (6): 679–82. https://doi.org/10.1038/s41592-022-01488-1.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, et al. 2020. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Miyazawa, Sanzo, and Robert L. Jernigan. 1985. "Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation." *Macromolecules* 18 (3): 534–52. https://doi.org/10.1021/ma00145a039.

Moen, Erick, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. 2019. "Deep Learning for Cellular Image Analysis." *Nature Methods* 16 (12): 1233–46. https://doi.org/10.1038/s41592-019-0403-1.

Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data." *Machine Learning* 52 (1-2): 91–118. https://doi.org/10.1023/a:1023949509487.

Moody, Edmund R. R., Sandra Álvarez-Carretero, Tara A. Mahendrarajah, et al. 2024. "The Nature of the Last Universal Common Ancestor and Its Impact on the Early Earth System." *Nature Ecology & Evolution* 8 (9): 1654–66. https://doi.org/10.1038/s41559-024-02461-1.

Moore, Josh, Chris Allan, Sébastien Besson, et al. 2021. "OME-NGFF: A Next-Generation File Format for Expanding Bioimaging Data-Access Strategies." *Nature Methods* 18 (12): 1496–98. https://doi.org/10.1038/s41592-021-01326-w.

Moore, Josh, Daniela Basurto-Lozada, Sébastien Besson, et al. 2023. "OME-Zarr: A Cloud-Optimized Bioimaging File Format with International Community Support." *Histochemistry and Cell Biology* 160 (3): 223–51. https://doi.org/10.1007/s00418-023-02209-1.

Morales, Joannella, Shashikant Pujar, Jane E. Loveland, et al. 2022. "A Joint NCBI and EMBL-EBI Transcript Set for Clinical Genomics and Research." *Nature* 604 (7905): 310–15. https://doi.org/10.1038/s41586-022-04558-8.

Morcos, Faruck, Andrea Pagnani, Bryan Lunt, et al. 2011. "Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts Across Many Protein Families." *Proceedings of the National Academy of Sciences* 108 (49). https://doi.org/10.1073/pnas.1111471108.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Moult, John, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. 1995. "A Large-Scale Experiment to Assess Protein Structure Prediction Methods." *Proteins: Structure, Function, and Bioinformatics* 23 (3). https://doi.org/10.1002/prot.340230303.

Mouse Genome Sequencing Consortium. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62. https://doi.org/10.1038/nature01262.

Moutakanni, Théo, Camille Couprie, Seungeun Yi, et al. 2025. "Cell-DINO: Self-Supervised Image-Based Embeddings for Cell Fluorescent Microscopy." *PLOS Computational Biology* 21 (12): e1013828. https://doi.org/10.1371/journal.pcbi.1013828.

Mudge, Jonathan M., Jorge Ruiz-Orera, John R. Prensner, et al. 2022. "Standardized Annotation of

Translated Open Reading Frames." *Nature Biotechnology* 40 (7): 994–99. https://doi.org/10.1038/s41587-022-01369-0.

Muegge, Ingo, and Yvonne C. Martin. 1999. "A General and Fast Scoring Function for Protein–ligand Interactions: A Simplified Potential Approach." *Journal of Medicinal Chemistry* 42 (5): 791–804. https://doi.org/10.1021/jm980536j.

Mund, Andreas, Fabian Coscia, András Kriston, et al. 2022. "Deep Visual Proteomics Defines Single-Cell Identity and Heterogeneity." *Nature Biotechnology* 40 (8): 1231–40. https://doi.org/10.1038/s41587-022-01302-5.

Murphy, William J., Denis M. Larkin, Annelie Everts-van der Wind, et al. 2005. "Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps." *Science* 309 (5734): 613–17. https://doi.org/10.1126/science.1111387.

Murzin, Alexey G., Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. 1995. "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures." *Journal of Molecular Biology* 247 (4): 536–40. https://doi.org/10.1016/s0022-2836(05)80134-2.

Mushegian, Arcady R., and Eugene V. Koonin. 1996. "Gene Order Is Not Conserved in Bacterial Evolution." *Trends in Genetics* 12 (8): 289–90. https://doi.org/10.1016/0168-9525(96)20006-x.

Myers, Eugene W. 2005. "The Fragment Assembly String Graph." *Bioinformatics* 21 (suppl_2): ii79–85. https://doi.org/10.1093/bioinformatics/bti1114.

Myers, Eugene W., and Webb Miller. 1988. "Optimal Alignments in Linear Space." *Bioinformatics* 4 (1): 11–17. https://doi.org/10.1093/bioinformatics/4.1.11.

Myers, Eugene W., Granger G. Sutton, Art L. Delcher, et al. 2000. "A Whole-Genome Assembly of Drosophila." *Science* 287 (5461): 2196–204. https://doi.org/10.1126/science.287.5461.2196.

Myers, Simon, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. 2005. "A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome." *Science* 310 (5746): 321–24. https://doi.org/10.1126/science.1117196.

Nadeau, J H, and B A Taylor. 1984. "Lengths of Chromosomal Segments Conserved Since Divergence of Man and Mouse." *Proceedings of the National Academy of Sciences* 81 (3): 814–18. https://doi.org/10.1073/pnas.81.3.814.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, et al. 2008. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science* 320 (5881): 1344–49. https://doi.org/10.1126/science.1158441.

Nagarajan, Niranjan, and Mihai Pop. 2013. "Sequence Assembly Demystified." *Nature Reviews Genetics* 14 (3): 157–67. https://doi.org/10.1038/nrg3367.

Nakabachi, Atsushi, Atsushi Yamashita, Hidehiro Toh, et al. 2006. "The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella." *Science* 314 (5797): 267–67. https://doi.org/10.1126/science.1134196.

Nakai, Kenta, and Paul Horton. 1999. "PSORT: A Program for Detecting Sorting Signals in Proteins

and Predicting Their Subcellular Localization." *Trends in Biochemical Sciences* 24 (1): 34–35. https://doi.org/10.1016/s0968-0004(98)01336-x.

Nam, Anna S., Ronan Chaligne, and Dan A. Landau. 2020. "Integrating Genetic and Non-Genetic Determinants of Cancer Evolution by Single-Cell Multi-Omics." *Nature Reviews Genetics* 22 (1): 3–18. https://doi.org/10.1038/s41576-020-0265-5.

Navin, Nicholas E. 2015. "The First Five Years of Single-Cell Cancer Genomics and Beyond." *Genome Research* 25 (10): 1499–507. https://doi.org/10.1101/gr.191098.115.

Navin, Nicholas, Jude Kendall, Jennifer Troge, et al. 2011. "Tumour Evolution Inferred by Single-Cell Sequencing." *Nature* 472 (7341): 90–94. https://doi.org/10.1038/nature09807.

Nawrocki, Eric P., and Sean R. Eddy. 2013. "Infernal 1.1: 100-Fold Faster RNA Homology Searches." *Bioinformatics* 29 (22): 2933–35. https://doi.org/10.1093/bioinformatics/btt509.

Nawrocki, Eric P., Diana L. Kolbe, and Sean R. Eddy. 2009. "Infernal 1.0: Inference of RNA Alignments." *Bioinformatics* 25 (10): 1335–37. https://doi.org/10.1093/bioinformatics/btp157.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, et al. 2020. "A Genomic Catalog of Earth's Microbiomes." *Nature Biotechnology* 39 (4): 499–509. https://doi.org/10.1038/s41587-020-0718-6.

Necci, Marco, Damiano Piovesan, Md Tamjidul Hoque, et al. 2021. "Critical Assessment of Protein Intrinsic Disorder Prediction." *Nature Methods* 18 (5): 472–81. https://doi.org/10.1038/s41592-021-01117-3.

Needleman, Saul B., and Christian D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48 (3): 443–53. https://doi.org/10.1016/0022-2836(70)90057-4.

Nehrt, Nathan L., Wyatt T. Clark, Predrag Radivojac, and Matthew W. Hahn. 2011. "Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals." *PLoS Computational Biology* 7 (6): e1002073. https://doi.org/10.1371/journal.pcbi.1002073.

Nei, Masatoshi, and Takashi Gojobori. 1986. "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions." *Molecular Biology and Evolution*, ahead of print. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

Nelson, Benjamin R., Catherine A. Makarewich, Douglas M. Anderson, et al. 2016. "A Peptide Encoded by a Transcript Annotated as Long Noncoding RNA Enhances SERCA Activity in Muscle." *Science* 351 (6270): 271–75. https://doi.org/10.1126/science.aad4076.

Nesvizhskii, Alexey I. 2010. "A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics." *Journal of Proteomics* 73 (11): 2092–123. https://doi.org/10.1016/j.jprot.2010.08.009.

Nesvizhskii, Alexey I., and Ruedi Aebersold. 2005. "Interpretation of Shotgun Proteomic Data." *Molecular & Cellular Proteomics* 4 (10): 1419–40. https://doi.org/10.1074/mcp.r500012-mcp200.

Nesvizhskii, Alexey I., Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. "A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry." *Analytical Chemistry* 75 (17): 4646–58.

https://doi.org/10.1021/ac0341261.

Newman, Aaron M, Chih Long Liu, Michael R Green, et al. 2015. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles." *Nature Methods* 12 (5): 453–57. https://doi.org/10.1038/nmeth.3337.

Ng, P. C. 2003. "SIFT: Predicting Amino Acid Changes That Affect Protein Function." *Nucleic Acids Research* 31 (13): 3812–14. https://doi.org/10.1093/nar/gkg509.

Ng, Sarah B, Kati J Buckingham, Choli Lee, et al. 2009. "Exome Sequencing Identifies the Cause of a Mendelian Disorder." *Nature Genetics* 42 (1): 30–35. https://doi.org/10.1038/ng.499.

Nguyen, Eric, Michael Poli, Matthew G. Durrant, et al. 2024. "Sequence Modeling and Design from Molecular to Genome Scale with Evo." *Science* 386 (6723). https://doi.org/10.1126/science.ado9336.

Nguyen, Lan Huong, and Susan Holmes. 2019. "Ten Quick Tips for Effective Dimensionality Reduction." *PLOS Computational Biology* 15 (6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907.

Nickell, Stephan, Christine Kofler, Andrew P. Leis, and Wolfgang Baumeister. 2006. "A Visual Approach to Proteomics." *Nature Reviews Molecular Cell Biology* 7 (3): 225–30. https://doi.org/10.1038/nrm1861.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. "Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites." *Protein Engineering Design and Selection* 10 (1): 1–6. https://doi.org/10.1093/protein/10.1.1.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. "Genotype and SNP Calling from Next-Generation Sequencing Data." *Nature Reviews Genetics* 12 (6): 443–51. https://doi.org/10.1038/nrg2986.

Nielsen, Rasmus, and Ziheng Yang. 1998. "Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene." *Genetics* 148 (3): 929–36. https://doi.org/10.1093/genetics/148.3.929.

Nieuwenhuis, Sander, Birte U Forstmann, and Eric-Jan Wagenmakers. 2011. "Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance." *Nature Neuroscience* 14 (9): 1105–7. https://doi.org/10.1038/nn.2886.

Nimwegen, Erik van. 2003. "Scaling Laws in the Functional Content of Genomes." *Trends in Genetics* 19 (9): 479–84. https://doi.org/10.1016/s0168-9525(03)00203-8.

Noble, William S. 2006. "What Is a Support Vector Machine?" *Nature Biotechnology* 24 (12): 1565–67. https://doi.org/10.1038/nbt1206-1565.

Noble, William S. 2009. "How Does Multiple Testing Correction Work?" *Nature Biotechnology* 27 (12): 1135–37. https://doi.org/10.1038/nbt1209-1135.

Noble, William Stafford, and Michael J. MacCoss. 2012. "Computational and Statistical Analysis of Protein Mass Spectrometry Data." *PLoS Computational Biology* 8 (1): e1002296. https://doi.org/10.1371/journal.pcbi.1002296.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the x-Inactivation Centre." *Nature* 485 (7398): 381–85. https://doi.org/10.1038/nature11

049.

Notin, Pascal, Aaron W. Kollasch, Daniel Ritter, et al. 2023. "ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design." *Advances in Neural Information Processing Systems* 36. https://proteingym.org.

Notredame, Cédric. 2007. "Recent Evolutions of Multiple Sequence Alignment Algorithms." *PLoS Computational Biology* 3 (8): e123. https://doi.org/10.1371/journal.pcbi.0030123.

Notredame, Cédric, Desmond G Higgins, and Jaap Heringa. 2000. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302 (1): 205–17. https://doi.org/10.1006/jmbi.2000.4042.

Novakovsky, Gherman, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. 2022. "Obtaining Genetics Insights from Deep Learning via Explainable Artificial Intelligence." *Nature Reviews Genetics* 24 (2): 125–37. https://doi.org/10.1038/s41576-022-00532-2.

Nowell, Peter C. 1976. "The Clonal Evolution of Tumor Cell Populations: Acquired Genetic Lability Permits Stepwise Selection of Variant Sublines and Underlies Tumor Progression." *Science* 194 (4260): 23–28. https://doi.org/10.1126/science.959840.

Nurk, Sergey, Sergey Koren, Arang Rhie, et al. 2022. "The Complete Sequence of a Human Genome." *Science* 376 (6588): 44–53. https://doi.org/10.1126/science.abj6987.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34. https://doi.org/10.1101/gr.213959.116.

Nussinov, R, and A B Jacobson. 1980. "Fast Algorithm for Predicting the Secondary Structure of Single-Stranded RNA." *Proceedings of the National Academy of Sciences* 77 (11): 6309–13. https://doi.org/10.1073/pnas.77.11.6309.

Nüsslein-Volhard, Christiane, and Eric Wieschaus. 1980. "Mutations Affecting Segment Number and Polarity in Drosophila." *Nature* 287 (5785): 795–801. https://doi.org/10.1038/287795a0.

Nuzzo, Regina. 2014. "Scientific Method: Statistical Errors." *Nature* 506 (7487): 150–52. https://doi.org/10.1038/506150a.

O'Farrell, PH. 1975. "High Resolution Two-Dimensional Electrophoresis of Proteins." *Journal of Biological Chemistry* 250 (10): 4007–21. https://doi.org/10.1016/s0021-9258(19)41496-8.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, et al. 2015. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1): D733–45. https://doi.org/10.1093/nar/gkv1189.

O'Reilly, Francis J., and Juri Rappsilber. 2018. "Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology." *Nature Structural & Molecular Biology* 25 (11): 1000–1008. https://doi.org/10.1038/s41594-018-0147-0.

O'Reilly, Francis J., Liang Xue, Andrea Graziadei, et al. 2020. "In-Cell Architecture of an Actively Transcribing-Translating Expressome." *Science* 369 (6503): 554–57. https://doi.org/10.1126/science.

abb3758.

Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. "Lateral Gene Transfer and the Nature of Bacterial Innovation." *Nature* 405 (6784): 299–304. https://doi.org/10.1038/35012500.

Ochoa, David, Andrew F. Jarnuczak, Cristina Viéitez, et al. 2019. "The Functional Landscape of the Human Phosphoproteome." *Nature Biotechnology* 38 (3): 365–73. https://doi.org/10.1038/s41587-019-0344-3.

Ofer, Dan, Nadav Brandes, and Michal Linial. 2021. "The Language of Proteins: NLP, Machine Learning & Protein Sequences." *Computational and Structural Biotechnology Journal* 19: 1750–58. https://doi.org/10.1016/j.csbj.2021.03.022.

Ogris, Christoph, Dimitri Guala, Thomas Helleday, and Erik L. L. Sonnhammer. 2016. "A Novel Method for Crosstalk Analysis of Biological Networks: Improving Accuracy of Pathway Annotation." *Nucleic Acids Research* 45 (2): e8–8. https://doi.org/10.1093/nar/gkw849.

Ohno, S., U. Wolf, and N. B. Atkin. 1968. "Evolution from Fish to Mammals by Gene Duplication." *Hereditas* 59: 169–87.

Ohta, Tomoko. 1973. "Slightly Deleterious Mutant Substitutions in Evolution." *Nature* 246 (5428): 96–98. https://doi.org/10.1038/246096a0.

Oliveira, Saulo de, and Charlotte Deane. 2017. "Co-Evolution Techniques Are Reshaping the Way We Do Structural Bioinformatics." *F1000Research* 6: 1224. https://doi.org/10.12688/f1000research.11543.1.

Oliver, Stephen G., Antonia Lock, Midori A. Harris, Paul Nurse, and Valerie Wood. 2016. "Model Organism Databases: Essential Resources That Need the Support of Both Funders and Users." *BMC Biology* 14 (1). https://doi.org/10.1186/s12915-016-0276-z.

Olsen, Jesper V., Blagoy Blagoev, Florian Gnad, et al. 2006. "Global, in Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks." *Cell* 127 (3): 635–48. https://doi.org/10.1016/j.cell.2006.09.026.

Omenn, Gilbert S., Lydie Lane, Christopher M. Overall, et al. 2024. "The 2023 Report on the Proteome from the HUPO Human Proteome Project." *Journal of Proteome Research* 23 (2): 532–49. https://doi.org/10.1021/acs.jproteome.3c00591.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, et al. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0997-x.

Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, et al. 2002. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics." *Molecular & Cellular Proteomics* 1 (5): 376–86. https://doi.org/10.1074/mcp.m200025-mcp200.

Ontiveros-Palacios, Nancy, Emma Cooke, Eric P Nawrocki, et al. 2024. "Rfam 15: RNA Families Database in 2025." *Nucleic Acids Research* 53 (D1): D258–67. https://doi.org/10.1093/nar/gkae1023.

OpenDDE contributors. 2026. *OpenDDE: Open Drug Discovery Engine*. GitHub repository. https://github.com/contact-ajmal/opendde.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, et al. 2023. *DINOv2: Learning Robust Visual Features Without Supervision*. https://doi.org/10.48550/arXiv.2304.07193.

Orgel, Leslie E. 2004. "Prebiotic Chemistry and the Origin of the RNA World." *Critical Reviews in Biochemistry and Molecular Biology* 39 (2): 99–123. https://doi.org/10.1080/10409230490460765.

Oró, J. 1960. "Synthesis of Adenine from Ammonium Cyanide." *Biochemical and Biophysical Research Communications* 2 (6): 407–12. https://doi.org/10.1016/0006-291x(60)90138-8.

Orth, Jeffrey D, Ines Thiele, and Bernhard Ø Palsson. 2010. "What Is Flux Balance Analysis?" *Nature Biotechnology* 28 (3): 245–48. https://doi.org/10.1038/nbt.1614.

Oshlack, Alicia, Mark D Robinson, and Matthew D Young. 2010. "From RNA-Seq Reads to Differential Expression Results." *Genome Biology* 11 (12): 220. https://doi.org/10.1186/gb-2010-11-12-220.

Oughtred, Rose, Jennifer Rust, Christie Chang, et al. 2020. "The BioGRID Database: A Comprehensive Biomedical Resource of Curated Protein, Genetic, and Chemical Interactions." *Protein Science* 30 (1): 187–200. https://doi.org/10.1002/pro.3978.

Ounkomol, Chawin, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. 2018. "Label-Free Prediction of Three-Dimensional Fluorescence Images from Transmitted-Light Microscopy." *Nature Methods* 15 (11): 917–20. https://doi.org/10.1038/s41592-018-0111-2.

Ouyang, Wei, Fynn Beuttenmueller, Estibaliz Gómez-de-Mariscal, et al. 2022. "BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2022.06.07.495102.

Ouyang, Wei, Casper F. Winsnes, Martin Hjelmare, et al. 2019. "Analysis of the Human Protein Atlas Image Classification Competition." *Nature Methods* 16 (12): 1254–61. https://doi.org/10.1038/s41592-019-0658-6.

Ovchinnikov, Sergey, Hetunandan Kamisetty, and David Baker. 2014. "Robust and Accurate Prediction of Residue–Residue Interactions Across Protein Interfaces Using Evolutionary Information." *eLife* 3. https://doi.org/10.7554/elife.02030.

Ovchinnikov, Sergey, Hahnbeom Park, Neha Varghese, et al. 2017. "Protein Structure Determination Using Metagenome Sequence Data." *Science* 355 (6322): 294–98. https://doi.org/10.1126/science.aah4043.

Overbeek, Ross, Michael Fonstein, Mark D'Souza, Gordon D. Pusch, and Natalia Maltsev. 1999. "The Use of Gene Clusters to Infer Functional Coupling." *Proceedings of the National Academy of Sciences* 96 (6): 2896–901. https://doi.org/10.1073/pnas.96.6.2896.

Oza, Vishal H., Jordan H. Whitlock, Elizabeth J. Wilk, et al. 2023. "Ten Simple Rules for Using Public Biological Data for Your Research." *PLOS Computational Biology* 19 (1): e1010749. https://doi.org/10.1371/journal.pcbi.1010749.

Pace, Norman R. 1997. "A Molecular View of Microbial Diversity and the Biosphere." *Science* 276 (5313): 734–40. https://doi.org/10.1126/science.276.5313.734.

Pacesa, Martin, Lennart Nickel, Christian Schellhaas, et al. 2025. "One-Shot Design of Functional Protein

Binders with BindCraft." *Nature* 646 (8084): 483–92. https://doi.org/10.1038/s41586-025-09429-6.

Pachitariu, Marius, and Carsen Stringer. 2022. "Cellpose 2.0: How to Train Your Own Model." *Nature Methods* 19 (12): 1634–41. https://doi.org/10.1038/s41592-022-01663-4.

Paggi, Joseph M., Ayush Pandit, and Ron O. Dror. 2024. "The Art and Science of Molecular Docking." *Annual Review of Biochemistry* 93 (1): 389–410. https://doi.org/10.1146/annurev-biochem-030222-120000.

Pal, Koustav, Mattia Forcato, and Francesco Ferrari. 2018. "Hi-c Analysis: From Data Generation to Integration." *Biophysical Reviews* 11 (1): 67–78. https://doi.org/10.1007/s12551-018-0489-1.

Palazzo, Alexander F., and T. Ryan Gregory. 2014. "The Case for Junk DNA." *PLoS Genetics* 10 (5): e1004351. https://doi.org/10.1371/journal.pgen.1004351.

Palazzo, Alexander F., and Eliza S. Lee. 2015. "Non-Coding RNA: What Is Functional and What Is Junk?" *Frontiers in Genetics* 6. https://doi.org/10.3389/fgene.2015.00002.

Palmer, Jeffrey D., and Laura A. Herbon. 1988. "Plant Mitochondrial DNA Evolved Rapidly in Structure, but Slowly in Sequence." *Journal of Molecular Evolution* 28 (1-2): 87–97. https://doi.org/10.1007/bf02143500.

Pan, Qun, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. 2008. "Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing." *Nature Genetics* 40 (12): 1413–15. https://doi.org/10.1038/ng.259.

Pan, Xingjie, and Tanja Kortemme. 2021. "Recent Advances in de Novo Protein Design: Principles, Methods, and Applications." *Journal of Biological Chemistry* 296: 100558. https://doi.org/10.1016/j.jbc.2021.100558.

Panneerselvam, Kalpana, Pablo Porras, Noemí del-Toro, et al. 2024. "IntAct Database for Accessing IMEx's Contextual Metadata of Molecular Interactions." *Current Protocols* 4 (10). https://doi.org/10.1002/cpz1.70018.

Papp, Balázs, Csaba Pál, and Laurence D. Hurst. 2003. "Dosage Sensitivity and the Evolution of Gene Families in Yeast." *Nature* 424 (6945): 194–97. https://doi.org/10.1038/nature01771.

Pardo-Palacios, Francisco J., Dingjie Wang, Fairlie Reese, et al. 2024. "Systematic Assessment of Long-Read RNA-Seq Methods for Transcript Identification and Quantification." *Nature Methods* 21 (7): 1349–63. https://doi.org/10.1038/s41592-024-02298-3.

Park, Jiwoon, Junbum Kim, Tyler Lewy, et al. 2022. "Spatial Omics Technologies at Multimodal and Single Cell/Subcellular Level." *Genome Biology* 23 (1). https://doi.org/10.1186/s13059-022-02824-6.

Park, Jong, Kevin Karplus, Christian Barrett, et al. 1998. "Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods." *Journal of Molecular Biology* 284 (4): 1201–10. https://doi.org/10.1006/jmbi.1998.2221.

Park, Yungki, and Edward M Marcotte. 2012. "Flaws in Evaluation Schemes for Pair-Input Computational Predictions." *Nature Methods* 9 (12): 1134–36. https://doi.org/10.1038/nmeth.2259.

Parker, Eric T., Henderson J. Cleaves, Jason P. Dworkin, et al. 2011. "Primordial Synthesis of Amines and Amino Acids in a 1958 Miller h 2 s-Rich Spark Discharge Experiment." *Proceedings of the National Academy of Sciences* 108 (14): 5526–31. https://doi.org/10.1073/pnas.1019191108.

Parker, Joe, Georgia Tsagkogeorga, James A. Cotton, et al. 2013. "Genome-Wide Signatures of Convergent Evolution in Echolocating Mammals." *Nature* 502 (7470): 228–31. https://doi.org/10.1038/nature12511.

Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2020. "A Complete Domain-to-Species Taxonomy for Bacteria and Archaea." *Nature Biotechnology* 38 (9): 1079–86. https://doi.org/10.1038/s41587-020-0501-8.

Parks, Donovan H, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2021. "GTDB: An Ongoing Census of Bacterial and Archaeal Diversity Through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy." *Nucleic Acids Research* 50 (D1): D785–94. https://doi.org/10.1093/nar/gkab776.

Parks, Donovan H, Maria Chuvochina, David W Waite, et al. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." *Nature Biotechnology* 36 (10): 996–1004. https://doi.org/10.1038/nbt.4229.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." *Cell* 176 (3): 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

Passaro, Saro, Gabriele Corso, Jeremy Wohlwend, et al. 2025. "Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.06.14.659707.

Patel, Anoop P., Itay Tirosh, John J. Trombetta, et al. 2014. "Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma." *Science* 344 (6190): 1396–401. https://doi.org/10.1126/science.1254257.

Paten, Benedict, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. 2008. "Enredo and Pecan: Genome-Wide Mammalian Consistency-Based Multiple Alignment with Paralogs." *Genome Research* 18 (11): 1814–28. https://doi.org/10.1101/gr.076554.108.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Patterson, Murray, Tobias Marschall, Nadia Pisanti, et al. 2015. "WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads." *Journal of Computational Biology* 22 (6): 498–509. https://doi.org/10.1089/cmb.2014.0157.

Patterson, Nick, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. 2006. "Genetic Evidence for Complex Speciation of Humans and Chimpanzees." *Nature* 441 (7097): 1103–8. https:

//doi.org/10.1038/nature04789.

Pauling, Linus, Robert B. Corey, and H. R. Branson. 1951. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain." *Proceedings of the National Academy of Sciences* 37 (4): 205–11. https://doi.org/10.1073/pnas.37.4.205.

Pauling, Linus, and Emile Zuckerkandl. 1963. "Chemical Paleogenetics: Molecular 'Restoration Studies' of Extinct Forms of Life." *Acta Chemica Scandinavica* 17: S9–16.

Paysan-Lafosse, Typhaine, Antonina Andreeva, Matthias Blum, et al. 2024. "The Pfam Protein Families Database: Embracing AI/ML." *Nucleic Acids Research* 53 (D1): D523–34. https://doi.org/10.1093/nar/gkae997.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, et al. 2022. "InterPro in 2022." *Nucleic Acids Research* 51 (D1): D418–27. https://doi.org/10.1093/nar/gkac993.

Pe'er, Itsik, Roman Yelensky, David Altshuler, and Mark J. Daly. 2008. "Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants." *Genetic Epidemiology* 32 (4): 381–85. https://doi.org/10.1002/gepi.20303.

Pearce, James D, Sara E Simmonds, Gita Mahmoudabadi, et al. 2025. "A Cross-Species Generative Cell Atlas Across 1.5 Billion Years of Evolution: The TranscriptFormer Single-Cell Model." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2025.04.25.650731.

Pearson, W R, and D J Lipman. 1988. "Improved Tools for Biological Sequence Comparison." *Proceedings of the National Academy of Sciences* 85 (8): 2444–48. https://doi.org/10.1073/pnas.85.8.2444.

Pearson, William R. 1998. "Empirical Statistical Estimates for Sequence Similarity Searches." *Journal of Molecular Biology* 276 (1): 71–84. https://doi.org/10.1006/jmbi.1997.1525.

Pearson, William R. 1995. "Comparison of Methods for Searching Protein Sequence Databases." *Protein Science* 4 (6): 1145–60. https://doi.org/10.1002/pro.5560040613.

Pearson, William R. 2013. "An Introduction to Sequence Similarity ('Homology') Searching." *Current Protocols in Bioinformatics* 42 (1). https://doi.org/10.1002/0471250953.bi0301s42.

Peck, Ariana, Joshua Hutchings, Jonathan Schwartz, et al. 2026. "Lessons Learned from a Kaggle Challenge for Particle Picking in Cryo-Electron Tomography." *Nature Methods* 23 (9): 1908–16. https://doi.org/10.1038/s41592-026-03198-4.

Pei, Baikang, Cristina Sisu, Adam Frankish, et al. 2012. "The GENCODE Pseudogene Resource." *Genome Biology* 13 (9). https://doi.org/10.1186/gb-2012-13-9-r51.

Peiffer, Daniel A., Jennie M. Le, Frank J. Steemers, et al. 2006. "High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping." *Genome Research* 16 (9): 1136–48. https://doi.org/10.1101/gr.5402306.

Pellegrini, Matteo, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. 1999. "Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles." *Proceedings of the National Academy of Sciences* 96 (8): 4285–88. https://doi.org/10.1073/pnas.96.8.4285.

Pellicer, Jaume, Oriane Hidalgo, Steven Dodsworth, and Ilia Leitch. 2018. "Genome Size Diversity and Its Impact on the Evolution of Land Plants." *Genes* 9 (2): 88. https://doi.org/10.3390/genes9020088.

Perez, Gerardo, Galt P Barber, Anna Benet-Pages, et al. 2024. "The UCSC Genome Browser Database: 2025 Update." *Nucleic Acids Research* 53 (D1): D1243–49. https://doi.org/10.1093/nar/gkae974.

Pérez-Cobas, Ana Elena, Laura Gomez-Valero, and Carmen Buchrieser. 2020. "Metagenomic Approaches in Microbial Ecology: An Update on Whole-Genome and Marker Gene Sequencing Analyses." *Microbial Genomics* 6 (8). https://doi.org/10.1099/mgen.0.000409.

Perez-Riverol, Yasset, Emanuele Alpi, Rui Wang, Henning Hermjakob, and Juan Antonio Vizcaíno. 2015. "Making Proteomics Data Accessible and Reusable: Current State of Proteomics Databases and Repositories." *PROTEOMICS* 15 (5-6): 930–50. https://doi.org/10.1002/pmic.201400302.

Perez-Riverol, Yasset, Jingwen Bai, Chakradhar Bandla, et al. 2021. "The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences." *Nucleic Acids Research* 50 (D1): D543–52. https://doi.org/10.1093/nar/gkab1038.

Perez-Riverol, Yasset, Chakradhar Bandla, Deepti J Kundu, et al. 2024. "The PRIDE Database at 20 Years: 2025 Update." *Nucleic Acids Research* 53 (D1): D543–53. https://doi.org/10.1093/nar/gkae1011.

Perez-Riverol, Yasset, Laurent Gatto, Rui Wang, et al. 2016. "Ten Simple Rules for Taking Advantage of Git and GitHub." *PLOS Computational Biology* 12 (7): e1004947. https://doi.org/10.1371/journal.pcbi.1004947.

Perez-Riverol, Yasset, Andrey Zorin, Gaurhari Dass, et al. 2019. "Quantifying the Impact of Public Omics Data." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-11461-w.

Perkins, David N., Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. 1999. "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data." *Electrophoresis* 20 (18): 3551–67. https://doi.org/10.1002/(sici)1522-2683(19991201)20:18%3C3551::aid-elps3551%3E3.0.co;2-2.

Perou, Charles M., Therese Sørlie, Michael B. Eisen, et al. 2000. "Molecular Portraits of Human Breast Tumours." *Nature* 406 (6797): 747–52. https://doi.org/10.1038/35021093.

Persson, Emma, Mateusz Kaduk, Sofia K. Forslund, and Erik L. L. Sonnhammer. 2019. "Domainoid: Domain-Oriented Orthology Inference." *BMC Bioinformatics* 20 (1). https://doi.org/10.1186/s12859-019-3137-2.

Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Pertea, Mihaela, and Steven L Salzberg. 2010. "Between a Chicken and a Grape: Estimating the Number of Human Genes." *Genome Biology* 11 (5): 206. https://doi.org/10.1186/gb-2010-11-5-206.

Pertea, Mihaela, Alaina Shumate, Geo Pertea, et al. 2018. "CHESS: A New Human Gene Catalog Curated from Thousands of Large-Scale RNA Sequencing Experiments Reveals Extensive Transcriptional Noise." *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-018-1590-2.

PERUTZ, M. F., M. G. ROSSMANN, ANN F. CULLIS, HILARY MUIRHEAD, GEORG WILL, and A. C. T. NORTH. 1960. "Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-å. Resolution, Obtained by x-Ray Analysis." *Nature* 185 (4711): 416–22. https://doi.org/10.1038/185416a0.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. "SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions." *Nature Methods* 8 (10): 785–86. https://doi.org/10.1038/nmeth.1701.

Peterson, Amelia C., Jason D. Russell, Derek J. Bailey, Michael S. Westphall, and Joshua J. Coon. 2012. "Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics." *Molecular & Cellular Proteomics* 11 (11): 1475–88. https://doi.org/10.1074/mcp.o112.020131.

Pevsner, Jonathan. 2015. *Bioinformatics and Functional Genomics*. 3rd ed. Wiley-Blackwell.

Pevzner, Pavel A., Haixu Tang, and Michael S. Waterman. 2001. "An Eulerian Path Approach to DNA Fragment Assembly." *Proceedings of the National Academy of Sciences* 98 (17): 9748–53. https://doi.org/10.1073/pnas.171285098.

Pevzner, Pavel, and Glenn Tesler. 2003. "Human and Mouse Genomic Sequences Reveal Extensive Breakpoint Reuse in Mammalian Evolution." *Proceedings of the National Academy of Sciences* 100 (13): 7672–77. https://doi.org/10.1073/pnas.1330369100.

Philippe, Hervé, Henner Brinkmann, Dennis V. Lavrov, et al. 2011. "Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough." *PLoS Biology* 9 (3): e1000602. https://doi.org/10.1371/journal.pbio.1000602.

Philippe, Nadège, Matthieu Legendre, Gabriel Doutre, et al. 2013. "Pandoraviruses: Amoeba Viruses with Genomes up to 2.5 Mb Reaching That of Parasitic Eukaryotes." *Science* 341 (6143): 281–86. https://doi.org/10.1126/science.1239181.

Phillips, Ben P., and Elizabeth A. Miller. 2021. "Membrane Protein Folding and Quality Control." *Current Opinion in Structural Biology* 69: 50–54. https://doi.org/10.1016/j.sbi.2021.03.003.

Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. "Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." *Nature Methods* 10 (11): 1096–98. https://doi.org/10.1038/nmeth.2639.

Pierce, Brian G., Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng. 2014. "ZDOCK Server: Interactive Docking Prediction of Protein–Protein Complexes and Symmetric Multimers." *Bioinformatics* 30 (12): 1771–73. https://doi.org/10.1093/bioinformatics/btu097.

Poliseno, Laura, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J. Haveman, and Pier Paolo Pandolfi. 2010. "A Coding-Independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology." *Nature* 465 (7301): 1033–38. https://doi.org/10.1038/nature09144.

Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. 2009. "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies." *Genome Research* 20 (1): 110–21. https://doi.org/10.1101/gr.097857.109.

Pollard, Katherine S., Sofie R. Salama, Nelle Lambert, et al. 2006. "An RNA Gene Expressed During Cortical Development Evolved Rapidly in Humans." *Nature* 443 (7108): 167–72. https://doi.org/10.1038/nature05113.

Pond, Sergei L. Kosakovsky, Simon D. W. Frost, and Spencer V. Muse. 2004. "HyPhy: Hypothesis Testing Using Phylogenies." *Bioinformatics* 21 (5): 676–79. https://doi.org/10.1093/bioinformatics/bti079.

Ponder, Jay W., and Frederic M. Richards. 1987. "Tertiary Templates for Proteins." *Journal of Molecular Biology* 193 (4): 775–91. https://doi.org/10.1016/0022-2836(87)90358-5.

Ponting, Chris P., and Wilfried Haerty. 2022. "Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review." *Annual Review of Genomics and Human Genetics* 23 (1): 153–72. https://doi.org/10.1146/annurev-genom-112921-123710.

Poplin, Ryan, Pi-Chuan Chang, David Alexander, et al. 2018. "A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks." *Nature Biotechnology* 36 (10): 983–87. https://doi.org/10.1038/nbt.4235.

Porubsky, David, Harriet Dashnow, Thomas A. Sasani, et al. 2025. "Human de Novo Mutation Rates from a Four-Generation Pedigree Reference." *Nature* 643 (8071): 427–36. https://doi.org/10.1038/s41586-025-08922-2.

Posada, D, and K A Crandall. 1998. "MODELTEST: Testing the Model of DNA Substitution." *Bioinformatics* 14 (9): 817–18. https://doi.org/10.1093/bioinformatics/14.9.817.

Powner, Matthew W., Béatrice Gerland, and John D. Sutherland. 2009. "Synthesis of Activated Pyrimidine Ribonucleotides in Prebiotically Plausible Conditions." *Nature* 459 (7244): 239–42. https://doi.org/10.1038/nature08013.

Pozzati, Gabriele, Wensi Zhu, Claudio Bassot, John Lamb, Petras Kundrotas, and Arne Elofsson. 2021. "Limits and Potential of Combined Folding and Docking." *Bioinformatics* 38 (4): 954–61. https://doi.org/10.1093/bioinformatics/btab760.

Prado-Martinez, Javier, Peter H. Sudmant, Jeffrey M. Kidd, et al. 2013. "Great Ape Genetic Diversity and Population History." *Nature* 499 (7459): 471–75. https://doi.org/10.1038/nature12228.

Pratapa, Aditya, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. 2020. "Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data." *Nature Methods* 17 (2): 147–54. https://doi.org/10.1038/s41592-019-0690-6.

Preiner, Martina, Silke Asche, Sidney Becker, et al. 2020. "The Future of Origin of Life Research: Bridging Decades-Old Divisions." *Life* 10 (3): 20. https://doi.org/10.3390/life10030020.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS ONE* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Pross, Addy, and Robert Pascal. 2013. "The Origin of Life: What We Know, What We Can Know and What We Will Never Know." *Open Biology* 3 (3): 120190. https://doi.org/10.1098/rsob.120190.

Prüfer, Kay, Fernando Racimo, Nick Patterson, et al. 2013. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49. https://doi.org/10.1038/nature12886.

Prusiner, Stanley B. 1982. "Novel Proteinaceous Infectious Particles Cause Scrapie." *Science* 216 (4542): 136–44. https://doi.org/10.1126/science.6801762.

Pryszcz, Leszek P., Jaime Huerta-Cepas, and Toni Gabaldón. 2010. "MetaPhOrs: Orthology and Paralogy Predictions from Multiple Phylogenetic Evidence Using a Consistency-Based Confidence Score." *Nucleic Acids Research* 39 (5): e32–32. https://doi.org/10.1093/nar/gkq953.

Puigbò, Pere, Yuri I. Wolf, and Eugene V. Koonin. 2019. "Genome-Wide Comparative Analysis of Phylogenetic Trees: The Prokaryotic Forest of Life." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_8.

Putnam, Nicholas H., Thomas Butts, David E. K. Ferrier, et al. 2008. "The Amphioxus Genome and the Evolution of the Chordate Karyotype." *Nature* 453 (7198): 1064–71. https://doi.org/10.1038/nature06967.

Qian, Jiang, Nicholas M. Luscombe, and Mark Gerstein. 2001. "Protein Family and Fold Occurrence in Genomes: Power-Law Behaviour and Evolutionary Model." *Journal of Molecular Biology* 313 (4): 673–81. https://doi.org/10.1006/jmbi.2001.5079.

Qian, Ning, and Terrence J. Sejnowski. 1988. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models." *Journal of Molecular Biology* 202 (4): 865–84. https://doi.org/10.1016/0022-2836(88)90564-5.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

Quackenbush, John. 2001. "Computational Analysis of Microarray Data." *Nature Reviews Genetics* 2 (6): 418–27. https://doi.org/10.1038/35076576.

Quince, Christopher, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. 2017. "Shotgun Metagenomics, from Sampling to Analysis." *Nature Biotechnology* 35 (9): 833–44. https://doi.org/10.1038/nbt.3935.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77 (2): 257–86. https://doi.org/10.1109/5.18626.

Radivojac, Predrag, Wyatt T Clark, Tal Ronnen Oron, et al. 2013. "A Large-Scale Evaluation of Computational Protein Function Prediction." *Nature Methods* 10 (3): 221–27. https://doi.org/10.1038/nmeth.2340.

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. "Stereochemistry of Polypeptide

Chain Configurations." *Journal of Molecular Biology* 7 (1): 95–99. https://doi.org/10.1016/s0022-2836(63)80023-6.

Rands, Chris M., Stephen Meader, Chris P. Ponting, and Gerton Lunter. 2014. "8.2." *PLoS Genetics* 10 (7): e1004525. https://doi.org/10.1371/journal.pgen.1004525.

Rannala, Bruce, and Ziheng Yang. 1996. "Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference." *Journal of Molecular Evolution* 43 (3): 304–11. https://doi.org/10.1007/bf02338839.

Rao, Roshan, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2020. "Transformer Protein Language Models Are Unsupervised Structure Learners." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2020.12.15.422761.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Raue, A., C. Kreutz, T. Maiwald, et al. 2009. "Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood." *Bioinformatics* 25 (15): 1923–29. https://doi.org/10.1093/bioinformatics/btp358.

Rautiainen, Mikko, Sergey Nurk, Brian P. Walenz, et al. 2023. "Telomere-to-Telomere Assembly of Diploid Chromosomes with Verkko." *Nature Biotechnology* 41 (10): 1474–82. https://doi.org/10.1038/s41587-023-01662-6.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. 2002. "Hierarchical Organization of Modularity in Metabolic Networks." *Science* 297 (5586): 1551–55. https://doi.org/10.1126/science.1073374.

Redelings, Benjamin D., and Marc A. Suchard. 2005. "Joint Bayesian Estimation of Alignment and Phylogeny." *Systematic Biology* 54 (3): 401–18. https://doi.org/10.1080/10635150590947041.

Reeck, Gerald R., Christoph de Haën, David C. Teller, et al. 1987. "'Homology' in Proteins and Nucleic Acids: A Terminology Muddle and a Way Out of It." *Cell* 50 (5): 667. https://doi.org/10.1016/0092-8674(87)90322-9.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, et al. 2017. "The Human Cell Atlas." *eLife* 6. https://doi.org/10.7554/elife.27041.

Reich, David, Richard E. Green, Martin Kircher, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60. https://doi.org/10.1038/nature09710.

Reis, Mario dos, and Ziheng Yang. 2019. "Bayesian Molecular Clock Dating Using Genome-Scale Datasets." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_10.

Remm, Maido, Christian E. V. Storm, and Erik L. L. Sonnhammer. 2001. "Automatic Clustering of Orthologs and in-Paralogs from Pairwise Species Comparisons." *Journal of Molecular Biology* 314 (5): 1041–52. https://doi.org/10.1006/jmbi.2000.5197.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2011. "HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment." *Nature Methods* 9 (2): 173–75. https://doi.org/10.1038/nmeth.1818.

Rennie, Martin Luke, and Michael R. Oliver. 2025. "Emerging Frontiers in Protein Structure Prediction Following the AlphaFold Revolution." *Journal of The Royal Society Interface* 22 (225). https://doi.org/10.1098/rsif.2024.0886.

Replogle, Joseph M., Reuben A. Saunders, Angela N. Pogson, et al. 2022. "Mapping Information-Rich Genotype-Phenotype Landscapes with Genome-Scale Perturb-Seq." *Cell* 185 (14): 2559–2575.e28. https://doi.org/10.1016/j.cell.2022.05.013.

Rhee, Hyun-Woo, Peng Zou, Namrata D. Udeshi, et al. 2013. "Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging." *Science* 339 (6125): 1328–31. https://doi.org/10.1126/science.1230593.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46. https://doi.org/10.1038/s41586-021-03451-0.

Rhie, Arang, Sergey Nurk, Monika Cechova, et al. 2023. "The Complete Sequence of a Human y Chromosome." *Nature* 621 (7978): 344–54. https://doi.org/10.1038/s41586-023-06457-y.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02134-9.

Rhodes, Daniel R, Scott A Tomlins, Sooryanarayana Varambally, et al. 2005. "Probabilistic Model of the Human Protein-Protein Interaction Network." *Nature Biotechnology* 23 (8): 951–59. https://doi.org/10.1038/nbt1103.

Ribeiro, António J M, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. 2017. "Mechanism and Catalytic Site Atlas (m-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites." *Nucleic Acids Research* 46 (D1): D618–23. https://doi.org/10.1093/nar/gkx1012.

Riccio, Cristian, Max L. Jansen, Linlin Guo, and Andreas Ziegler. 2024. "Variant Effect Predictors: A Systematic Review and Practical Guide." *Human Genetics* 143 (5): 625–34. https://doi.org/10.1007/s00439-024-02670-5.

Rice, Gavin, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. 2023. "TomoTwin: Generalized 3D Localization of Macromolecules in Cryo-Electron Tomograms with Structural Data Mining." *Nature Methods* 20 (6): 871–80. https://doi.org/10.1038/s41592-023-01878-z.

Richards, Alicia L, Manon Eckhardt, and Nevan J Krogan. 2021. "Mass Spectrometry-Based Protein–Protein Interaction Networks for the Study of Human Diseases." *Molecular Systems Biology* 17 (1). https://doi.org/10.15252/msb.20188792.

Richards, Sue, Nazneen Aziz, Sherri Bale, et al. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine* 17 (5):

405–24. https://doi.org/10.1038/gim.2015.30.

Richardson, Jane S. 1981. "The Anatomy and Taxonomy of Protein Structure." In *Advances in Protein Chemistry Volume 34*. Elsevier. https://doi.org/10.1016/s0065-3233(08)60520-3.

Richardson, Lorna, Ben Allen, Germana Baldi, et al. 2022. "MGnify: The Microbiome Sequence Data Analysis Resource in 2023." *Nucleic Acids Research* 51 (D1): D753–59. https://doi.org/10.1093/nar/gkac1080.

Riesselman, Adam J., John B. Ingraham, and Debora S. Marks. 2018. "Deep Generative Models of Genetic Variation Capture the Effects of Mutations." *Nature Methods* 15 (10): 816–22. https://doi.org/10.1038/s41592-018-0138-4.

Rigaut, Guillaume, Anna Shevchenko, Berthold Rutz, Matthias Wilm, Matthias Mann, and Bertrand Séraphin. 1999. "A Generic Protein Purification Method for Protein Complex Characterization and Proteome Exploration." *Nature Biotechnology* 17 (10): 1030–32. https://doi.org/10.1038/13732.

Rigden, Daniel J, and Xosé M Fernández. 2025. "The 2026 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection." *Nucleic Acids Research* 54 (D1): D1–9. https://doi.org/10.1093/nar/gkaf1427.

Rissin, David M, Cheuk W Kan, Todd G Campbell, et al. 2010. "Single-Molecule Enzyme-Linked Immunosorbent Assay Detects Serum Proteins at Subfemtomolar Concentrations." *Nature Biotechnology* 28 (6): 595–99. https://doi.org/10.1038/nbt.1641.

Rivas, Elena, Jody Clements, and Sean R Eddy. 2016. "A Statistical Test for Conserved RNA Structure Shows Lack of Evidence for Structure in lncRNAs." *Nature Methods* 14 (1): 45–48. https://doi.org/10.1038/nmeth.4066.

Rives, Alexander, Joshua Meier, Tom Sercu, et al. 2021. "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences." *Proceedings of the National Academy of Sciences* 118 (15). https://doi.org/10.1073/pnas.2016239118.

Robinson, Jonathan L., Pınar Kocabaş, Hao Wang, et al. 2020. "An Atlas of Human Metabolism." *Science Signaling* 13 (624). https://doi.org/10.1126/scisignal.aaz1482.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. "edgeR : A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, Mark D, and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3). https://doi.org/10.1186/gb-2010-11-3-r25.

Rocha, Eduardo P. C. 2008. "The Organization of the Bacterial Genome." *Annual Review of Genetics* 42 (1): 211–33. https://doi.org/10.1146/annurev.genet.42.110807.091653.

Rodriques, Samuel G., Robert R. Stickels, Aleksandrina Goeva, et al. 2019. "Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution." *Science* 363 (6434): 1463–67. https://doi.org/10.1126/science.aaw1219.

Roger, Andrew J., Sergio A. Muñoz-Gómez, and Ryoma Kamikawa. 2017. "The Origin and Diversification

of Mitochondria." *Current Biology* 27 (21): R1177–92. https://doi.org/10.1016/j.cub.2017.09.015.

Rokas, Antonis, Barry L. Williams, Nicole King, and Sean B. Carroll. 2003. "Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies." *Nature* 425 (6960): 798–804. https://doi.org/10.1038/nature02053.

Rolland, Thomas, Murat Taşan, Benoit Charloteaux, et al. 2014. "A Proteome-Scale Map of the Human Interactome Network." *Cell* 159 (5): 1212–26. https://doi.org/10.1016/j.cell.2014.10.050.

Ronaghi, Mostafa, Mathias Uhlén, and Pål Nyrén. 1998. "A Sequencing Method Based on Real-Time Pyrophosphate." *Science* 281 (5375): 363–65. https://doi.org/10.1126/science.281.5375.363.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, et al. 2012. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space." *Systematic Biology* 61 (3): 539–42. https://doi.org/10.1093/sysbio/sys029.

Rood, Jennifer E., Samantha Wynne, Lucia Robson, et al. 2024. "The Human Cell Atlas from a Cell Census to a Unified Foundation Model." *Nature* 637 (8048): 1065–71. https://doi.org/10.1038/s41586-024-08338-4.

Roohani, Yusuf H., Tony J. Hua, Po-Yuan Tung, et al. 2025. "Virtual Cell Challenge: Toward a Turing Test for the Virtual Cell." *Cell* 188 (13): 3370–74. https://doi.org/10.1016/j.cell.2025.06.008.

Roohani, Yusuf, Kexin Huang, and Jure Leskovec. 2023. "Predicting Transcriptional Outcomes of Novel Multigene Perturbations with GEARS." *Nature Biotechnology* 42 (6): 927–35. https://doi.org/10.1038/s41587-023-01905-6.

Rosenblatt, F. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408. https://doi.org/10.1037/h0042519.

Rosonovski, Summer, Maria Levchenko, Michele Ide-Smith, Lynne Faulk, Melissa Harrison, and Johanna McEntyre. 2023. "Searching and Evaluating Publications and Preprints Using Europe PMC." *Current Protocols* 3 (3). https://doi.org/10.1002/cpz1.694.

Ross, Philip L., Yulin N. Huang, Jason N. Marchese, et al. 2004. "Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents." *Molecular & Cellular Proteomics* 3 (12): 1154–69. https://doi.org/10.1074/mcp.m400129-mcp200.

Rossi, Andrea, Zacharias Kontarakis, Claudia Gerri, et al. 2015. "Genetic Compensation Induced by Deleterious Mutations but Not Gene Knockdowns." *Nature* 524 (7564): 230–33. https://doi.org/10.1038/nature14580.

Rost, Burkhard. 1997. "Protein Structures Sustain Evolutionary Drift." *Folding and Design* 2: S19–24. https://doi.org/10.1016/s1359-0278(97)00059-x.

Rost, Burkhard. 1999. "Twilight Zone of Protein Sequence Alignments." *Protein Engineering, Design and Selection* 12 (2): 85–94. https://doi.org/10.1093/protein/12.2.85.

Rost, Burkhard. 2002. "Enzyme Function Less Conserved Than Anticipated." *Journal of Molecular Biology* 318 (2): 595–608. https://doi.org/10.1016/s0022-2836(02)00016-5.

Rost, Burkhard, and Chris Sander. 1993. "Prediction of Protein Secondary Structure at Better Than 70." *Journal of Molecular Biology* 232 (2): 584–99. https://doi.org/10.1006/jmbi.1993.1413.

Rost, Burkhard, Chris Sander, and Reinhard Schneider. 1994. "Redefining the Goals of Protein Secondary Structure Prediction." *Journal of Molecular Biology* 235 (1): 13–26. https://doi.org/10.1016/s0022-2836(05)80007-5.

Röst, Hannes L, George Rosenberger, Pedro Navarro, et al. 2014. "OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data." *Nature Biotechnology* 32 (3): 219–23. https://doi.org/10.1038/nbt.2841.

Roth, Alexander CJ, Gaston H Gonnet, and Christophe Dessimoz. 2008. "Algorithm of OMA for Large-Scale Orthology Inference." *BMC Bioinformatics* 9 (1). https://doi.org/10.1186/1471-2105-9-518.

Rothfels, Karen, Marija Milacic, Lisa Matthews, et al. 2023. "Using the Reactome Database." *Current Protocols* 3 (4). https://doi.org/10.1002/cpz1.722.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rout, Michael P., and Andrej Sali. 2019. "Principles for Integrative Structural Biology Studies." *Cell* 177 (6): 1384–403. https://doi.org/10.1016/j.cell.2019.05.016.

Roux, Kyle J., Dae In Kim, Manfred Raida, and Brian Burke. 2012. "A Promiscuous Biotin Ligase Fusion Protein Identifies Proximal and Interacting Proteins in Mammalian Cells." *Journal of Cell Biology* 196 (6): 801–10. https://doi.org/10.1083/jcb.201112098.

Rual, Jean-François, Kavitha Venkatesan, Tong Hao, et al. 2005. "Towards a Proteome-Scale Map of the Human Protein–Protein Interaction Network." *Nature* 437 (7062): 1173–78. https://doi.org/10.1038/nature04209.

Rueden, Curtis T., Johannes Schindelin, Mark C. Hiner, et al. 2017. "ImageJ2: ImageJ for the Next Generation of Scientific Image Data." *BMC Bioinformatics* 18 (1). https://doi.org/10.1186/s12859-017-1934-z.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–36. https://doi.org/10.1038/323533a0.

Sagan, Lynn. 1967. "On the Origin of Mitosing Cells." *Journal of Theoretical Biology* 14 (3): 225–IN6. https://doi.org/10.1016/0022-5193(67)90079-3.

Sahakyan, Harutyun, Pascal Mutz, Victor Tobiasson, and Eugene V. Koonin. 2025. "Exploring the Protein Universe with Distant Similarity Detection Methods." *Protein Science* 35 (1). https://doi.org/10.1002/pro.70397.

Saito, Takaya, and Marc Rehmsmeier. 2015. "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE* 10 (3): e0118432.

https://doi.org/10.1371/journal.pone.0118432.

Saitou, Naruya, and Masatoshi Nei. 1987. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4 (4): 406–25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Šali, Andrej, and Tom L. Blundell. 1993. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* 234 (3): 779–815. https://doi.org/10.1006/jmbi.1993.1626.

Salzberg, Steven L. 2019. "Next-Generation Genome Annotation: We Still Struggle to Get It Right." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1715-2.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. "Ten Simple Rules for Reproducible Computational Research." *PLoS Computational Biology* 9 (10): e1003285. https://doi.org/10.1371/journal.pcbi.1003285.

Sanger, F., G. M. Air, B. G. Barrell, et al. 1977. "Nucleotide Sequence of Bacteriophage φX174 DNA." *Nature* 265 (5596): 687–95. https://doi.org/10.1038/265687a0.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Sanjuán, Rafael, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. 2010. "Viral Mutation Rates." *Journal of Virology* 84 (19): 9733–48. https://doi.org/10.1128/jvi.00694-10.

Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, et al. 2014. "The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans." *Nature* 507 (7492): 354–57. https://doi.org/10.1038/nature12961.

Sankoff, David. 1975. "Minimal Mutation Trees of Sequences." *SIAM Journal on Applied Mathematics* 28 (1): 35–42. https://doi.org/10.1137/0128004.

Sankoff, David. 1985. "Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems." *SIAM Journal on Applied Mathematics* 45 (5): 810–25. https://doi.org/10.1137/0145048.

Sapoval, Nicolae, Amirali Aghazadeh, Michael G. Nute, et al. 2022. "Current Progress and Open Challenges for Applying Deep Learning Across the Biosciences." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-29268-7.

Sarich, Vincent M., and Allan C. Wilson. 1973. "Generation Time and Genomic Evolution in Primates." *Science* 179 (4078): 1144–47. https://doi.org/10.1126/science.179.4078.1144.

Sarkans, Ugis, Wah Chiu, Lucy Collinson, et al. 2021. "REMBI: Recommended Metadata for Biological Images—Enabling Reuse of Microscopy Data in Biology." *Nature Methods* 18 (12): 1418–22. https://doi.org/10.1038/s41592-021-01166-8.

Satam, Heena, Kandarp Joshi, Upasana Mangrolia, et al. 2023. "Next-Generation Sequencing Technology: Current Trends and Advancements." *Biology* 12 (7): 997. https://doi.org/10.3390/biology12070997.

Savage, Sara R., and Bing Zhang. 2020. "Using Phosphoproteomics Data to Understand Cellular

Signaling: A Comprehensive Guide to Bioinformatics Resources." *Clinical Proteomics* 17 (1). https://doi.org/10.1186/s12014-020-09290-x.

Savitski, Mikhail M., Friedrich B. M. Reinhard, Holger Franken, et al. 2014. "Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome." *Science* 346 (6205). https://doi.org/10.1126/science.1255784.

Savitski, Mikhail M., Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. 2015. "A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets." *Molecular & Cellular Proteomics* 14 (9): 2394–404. https://doi.org/10.1074/mcp.m114.046995.

Sawaya, Michael R., Michael P. Hughes, Jose A. Rodriguez, Roland Riek, and David S. Eisenberg. 2021. "The Expanding Amyloid Family: Structure, Stability, Function, and Pathogenesis." *Cell* 184 (19): 4857–73. https://doi.org/10.1016/j.cell.2021.08.013.

Sawaya, Michael R., Shilpa Sambashivan, Rebecca Nelson, et al. 2007. "Atomic Structures of Amyloid Cross-β Spines Reveal Varied Steric Zippers." *Nature* 447 (7143): 453–57. https://doi.org/10.1038/nature05695.

Sayers, Eric W, Jeffrey Beck, Evan E Bolton, et al. 2024. "Database Resources of the National Center for Biotechnology Information in 2025." *Nucleic Acids Research* 53 (D1): D20–29. https://doi.org/10.1093/nar/gkae979.

Sberro, Hila, Brayon J. Fremin, Soumaya Zlitni, et al. 2019. "Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes." *Cell* 178 (5): 1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.

Scally, Aylwyn, and Richard Durbin. 2012. "Revising the Human Mutation Rate: Implications for Understanding Human Evolution." *Nature Reviews Genetics* 13 (10): 745–53. https://doi.org/10.1038/nrg3295.

Scally, Aylwyn, Julien Y. Dutheil, LaDeana W. Hillier, et al. 2012. "Insights into Hominid Evolution from the Gorilla Genome Sequence." *Nature* 483 (7388): 169–75. https://doi.org/10.1038/nature10842.

Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, et al. 2001. "Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-Based Statistics and Other Refinements." *Nucleic Acids Research* 29 (14): 2994–3005. https://doi.org/10.1093/nar/29.14.2994.

Schaid, Daniel J., Wenan Chen, and Nicholas B. Larson. 2018. "From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping." *Nature Reviews Genetics* 19 (8): 491–504. https://doi.org/10.1038/s41576-018-0016-z.

Schena, Mark, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270 (5235): 467–70. https://doi.org/10.1126/science.270.5235.467.

Scheres, Sjors H. W. 2012. "A Bayesian View on Cryo-EM Structure Determination." *Journal of Molecular Biology* 415 (2): 406–18. https://doi.org/10.1016/j.jmb.2011.11.010.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, et al. 2012. "Fiji: An Open-Source Platform for Biological-Image Analysis." *Nature Methods* 9 (7): 676–82. https://doi.org/10.1038/nmeth.2019.

Schleper, Christa, and Thiago Rodrigues-Oliveira. 2026. "Asgard Archaea: Have We Found Our Microbial Ancestors?" *The EMBO Journal* 45 (6): 1836–51. https://doi.org/10.1038/s44318-026-00719-x.

Schmidt, Uwe, Martin Weigert, Coleman Broaddus, and Gene Myers. 2018. "Cell Detection with Star-Convex Polygons." In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing. https://doi.org/10.1007/978-3-030-00934-2_30.

Schneider, Thomas D., and R.Michael Stephens. 1990. "Sequence Logos: A New Way to Display Consensus Sequences." *Nucleic Acids Research* 18 (20): 6097–100. https://doi.org/10.1093/nar/18.20.6097.

Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies." *PLoS Computational Biology* 5 (12): e1000605. https://doi.org/10.1371/journal.pcbi.1000605.

Schnoes, Alexandra M., David C. Ream, Alexander W. Thorman, Patricia C. Babbitt, and Iddo Friedberg. 2013. "Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space." *PLoS Computational Biology* 9 (5): e1003063. https://doi.org/10.1371/journal.pcbi.1003063.

Schoch, Conrad L, Stacy Ciufo, Mikhail Domrachev, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database* 2020. https://doi.org/10.1093/database/baaa062.

Schopper, Simone, Abdullah Kahraman, Pascal Leuenberger, et al. 2017. "Measuring Protein Structural Changes on a Proteome-Wide Scale Using Limited Proteolysis-Coupled Mass Spectrometry." *Nature Protocols* 12 (11): 2391–410. https://doi.org/10.1038/nprot.2017.100.

Schreiber, Fabian, and Erik L. L. Sonnhammer. 2013. "Hieranoid: Hierarchical Orthology Inference." *Journal of Molecular Biology* 425 (11): 2072–81. https://doi.org/10.1016/j.jmb.2013.02.018.

Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, et al. 2016. "How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?" *RNA* 22 (6): 839–51. https://doi.org/10.1261/rna.053959.115.

Schwanhäusser, Björn, Dorothea Busse, Na Li, et al. 2011. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473 (7347): 337–42. https://doi.org/10.1038/nature10098.

Schwartz, Russell, and Alejandro A. Schäffer. 2017. "The Evolution of Tumour Phylogenetics: Principles and Practice." *Nature Reviews Genetics* 18 (4): 213–29. https://doi.org/10.1038/nrg.2016.170.

Schwartz, Scott, W. James Kent, Arian Smit, et al. 2002. "Human–Mouse Alignments with BLASTZ." *Genome Research* 13 (1): 103–7. https://doi.org/10.1101/gr.809403.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, et al. 2017. "Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software." *Nature Methods* 14 (11): 1063–71. https://doi.org/10.1038/nmeth.4458.

Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, et al. 2018. "Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing." *Nature Methods* 15 (6): 461–68. https://doi.org/10.1038/s41592-018-0001-7.

Sela, Itamar, Yuri I. Wolf, and Eugene V. Koonin. 2016. "Theory of Prokaryotic Genome Evolution." *Proceedings of the National Academy of Sciences* 113 (41): 11399–407. https://doi.org/10.1073/pnas.1614083113.

Senior, Andrew W., Richard Evans, John Jumper, et al. 2020. "Improved Protein Structure Prediction Using Potentials from Deep Learning." *Nature* 577 (7792): 706–10. https://doi.org/10.1038/s41586-019-1923-7.

Shalem, Ophir, Neville E. Sanjana, Ella Hartenian, et al. 2014. "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells." *Science* 343 (6166): 84–87. https://doi.org/10.1126/science.1247005.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shapovalov, Maxim V., and Roland L. Dunbrack. 2011. "A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions." *Structure* 19 (6): 844–58. https://doi.org/10.1016/j.str.2011.03.019.

Sharan, Roded, Igor Ulitsky, and Ron Shamir. 2007. "Network-Based Prediction of Protein Function." *Molecular Systems Biology* 3 (1). https://doi.org/10.1038/msb4100129.

Shaw, David E., Paul Maragakis, Kresten Lindorff-Larsen, et al. 2010. "Atomic-Level Characterization of the Structural Dynamics of Proteins." *Science* 330 (6002): 341–46. https://doi.org/10.1126/science.1187409.

Shendure, Jay, Shankar Balasubramanian, George M. Church, et al. 2017. "DNA Sequencing at 40: Past, Present and Future." *Nature* 550 (7676): 345–53. https://doi.org/10.1038/nature24286.

Shen-Orr, Shai S., Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. "Network Motifs in the Transcriptional Regulation Network of Escherichia Coli." *Nature Genetics* 31 (1): 64–68. https://doi.org/10.1038/ng881.

Shimodaira, Hidetoshi. 2002. "An Approximately Unbiased Test of Phylogenetic Tree Selection." *Systematic Biology* 51 (3): 492–508. https://doi.org/10.1080/10635150290069913.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3). https://doi.org/10.1214/10-sts330.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, et al. 2005. "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes." *Genome Research* 15 (8): 1034–50. https://doi.org/10.1101/gr.3715005.

Sievers, Fabian, Andreas Wilm, David Dineen, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (1). https://doi.org/10.1038/msb.2011.75.

Sigrist, Christian J. A., Edouard de Castro, Lorenzo Cerutti, et al. 2012. "New and Continuing Developments at PROSITE." *Nucleic Acids Research* 41 (D1): D344–47. https://doi.org/10.1093/nar/gks1067.

Sillitoe, Ian, Nicola Bordin, Natalie Dawson, et al. 2020. "CATH: Increased Structural Coverage of Functional Space." *Nucleic Acids Research* 49 (D1): D266–73. https://doi.org/10.1093/nar/gkaa1079.

Simakov, Oleg, Jessen Bredeson, Kodiak Berkoff, et al. 2022. "Deeply Conserved Synteny and the Evolution of Metazoan Chromosomes." *Science Advances* 8 (5). https://doi.org/10.1126/sciadv.abi5884.

Simakov, Oleg, Ferdinand Marlétaz, Jia-Xing Yue, et al. 2020. "Deeply Conserved Synteny Resolves Early Events in Vertebrate Evolution." *Nature Ecology & Evolution* 4 (6): 820–30. https://doi.org/10.1038/s41559-020-1156-z.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Simon, R., M. D. Radmacher, K. Dobbin, and L. M. McShane. 2003. "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification." *JNCI Journal of the National Cancer Institute* 95 (1): 14–18. https://doi.org/10.1093/jnci/95.1.14.

Simons, Kim T., Charles Kooperberg, Enoch Huang, and David Baker. 1997. "Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions." *Journal of Molecular Biology* 268 (1): 209–25. https://doi.org/10.1006/jmbi.1997.0959.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. M. Jones, and İnanç Birol. 2009. "ABySS: A Parallel Assembler for Short Read Sequence Data." *Genome Research* 19 (6): 1117–23. https://doi.org/10.1101/gr.089532.108.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp. 2017. "Detecting DNA Cytosine Methylation Using Nanopore Sequencing." *Nature Methods* 14 (4): 407–10. https://doi.org/10.1038/nmeth.4184.

Singh, Amrit, Casey P Shannon, Benoît Gautier, et al. 2019. "DIABLO: An Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays." *Bioinformatics* 35 (17): 3055–62. https://doi.org/10.1093/bioinformatics/bty1054.

Sippl, Manfred J. 1990. "Calculation of Conformational Ensembles from Potentials of Mena Force." *Journal of Molecular Biology* 213 (4): 859–83. https://doi.org/10.1016/s0022-2836(05)80269-4.

Sirén, Jouni, Jean Monlong, Xian Chang, et al. 2021. "Pangenomics Enables Genotyping of Known Structural Variants in 5202 Diverse Genomes." *Science* 374 (6574). https://doi.org/10.1126/science.abg8871.

Sjölander, Kimmen, Kevin Karplus, Michael Brown, et al. 1996. "Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology." *Bioinformatics* 12 (4): 327–45. https://doi.org/10.1093/bioinformatics/12.4.327.

Škunca, Nives, Richard J. Roberts, and Martin Steffen. 2016. "Evaluating Computational Gene Ontology Annotations." In *The Gene Ontology Handbook*. Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_8.

Skwark, Marcin J., Daniele Raimondi, Mirco Michel, and Arne Elofsson. 2014. "Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns." *PLoS Computational Biology* 10 (11): e1003889. https://doi.org/10.1371/journal.pcbi.1003889.

Slavoff, Sarah A, Andrew J Mitchell, Adam G Schwaid, et al. 2012. "Peptidomic Discovery of Short Open Reading Frame–Encoded Peptides in Human Cells." *Nature Chemical Biology* 9 (1): 59–64. https://doi.org/10.1038/nchembio.1120.

Slon, Viviane, Fabrizio Mafessoni, Benjamin Vernot, et al. 2018. "The Genome of the Offspring of a Neanderthal Mother and a Denisovan Father." *Nature* 561 (7721): 113–16. https://doi.org/10.1038/s41586-018-0455-x.

Slonim, Donna K., and Itai Yanai. 2009. "Getting Started in Gene Expression Microarray Analysis." *PLoS Computational Biology* 5 (10): e1000543. https://doi.org/10.1371/journal.pcbi.1000543.

Smedley, Damian, Julius O B Jacobsen, Marten Jäger, et al. 2015. "Next-Generation Diagnostics and Disease-Gene Discovery with the Exomiser." *Nature Protocols* 10 (12): 2004–15. https://doi.org/10.1038/nprot.2015.124.

Smith, Gabriel Reuben, Carolina Bello, Lalasia Bialic-Murphy, et al. 2024. "Ten Simple Rules for Using Large Language Models in Science, Version 1.0." *PLOS Computational Biology* 20 (1): e1011767. https://doi.org/10.1371/journal.pcbi.1011767.

Smith, John Maynard, and John Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene." *Genetical Research* 23 (1): 23–35. https://doi.org/10.1017/s0016672300014634.

Smith, Lloyd M., Jane Z. Sanders, Robert J. Kaiser, et al. 1986. "Fluorescence Detection in Automated DNA Sequence Analysis." *Nature* 321 (6071): 674–79. https://doi.org/10.1038/321674a0.

Smith, T. F., and M. S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1–25. https://doi.org/10.2202/1544-6115.1027.

Snel, Berend, Peer Bork, and Martijn Huynen. 2000. "Genome Evolution." *Trends in Genetics* 16 (1): 9–11. https://doi.org/10.1016/s0168-9525(99)01924-1.

Snider, Jamie, Max Kotlyar, Punit Saraon, Zhong Yao, Igor Jurisica, and Igor Stagljar. 2015. "Fundamentals of Protein Interaction Network Mapping." *Molecular Systems Biology* 11 (12). https://doi.org/10.15252/msb.20156351.

Soding, J., A. Biegert, and A. N. Lupas. 2005. "The HHpred Interactive Server for Protein Homology Detection and Structure Prediction." *Nucleic Acids Research* 33 (Web Server): W244–48. https://doi.org/10.1093/nar/gki408.

Söding, Johannes. 2004. "Protein Homology Detection by HMM–HMM Comparison." *Bioinformatics* 21 (7): 951–60. https://doi.org/10.1093/bioinformatics/bti125.

Sollis, Elliot, Abayomi Mosaku, Ala Abid, et al. 2022. "The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource." *Nucleic Acids Research* 51 (D1): D977–85. https://doi.org/10.1093/nar/gkac1010.

Sonawane, Abhijeet R., Scott T. Weiss, Kimberly Glass, and Amitabh Sharma. 2019. "Network Medicine

in the Age of Biomedical Big Data." *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.00294.

Sondka, Zbyslaw, Nidhi Bindal Dhir, Denise Carvalho-Silva, et al. 2023. "COSMIC: A Curated Database of Somatic Variants and Clinical Data for Cancer." *Nucleic Acids Research* 52 (D1): D1210–17. https://doi.org/10.1093/nar/gkad986.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2016. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." *F1000Research* 4: 1521. https://doi.org/10.12688/f1000research.7563.2.

Sonnhammer, Erik L. L., and Richard Durbin. 1995. "A Dot-Matrix Program with Dynamic Threshold Control Suited for Genomic DNA and Protein Sequence Analysis." *Gene* 167 (1-2): GC1–10. https://doi.org/10.1016/0378-1119(95)00714-8.

Sonnhammer, Erik L. L., Sean R. Eddy, and Richard Durbin. 1997. "Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments." *Proteins: Structure, Function, and Genetics* 28 (3): 405–20. https://doi.org/10.1002/(sici)1097-0134(199707)28:3%3C405::aid-prot10%3E3.0.co;2-l.

Spang, Anja, Tara A Mahendrarajah, Pierre Offre, and Courtney W Stairs. 2022. "Evolving Perspective on the Origin and Diversification of Cellular Life and the Virosphere." *Genome Biology and Evolution* 14 (6). https://doi.org/10.1093/gbe/evac034.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, et al. 2015. "Complex Archaea That Bridge the Gap Between Prokaryotes and Eukaryotes." *Nature* 521 (7551): 173–79. https://doi.org/10.1038/nature14447.

Spudich, Giulietta M, and Xosé M Fernández-Suárez. 2010. "Touring Ensembl: A Practical Guide to Genome Browsing." *BMC Genomics* 11 (1): 295. https://doi.org/10.1186/1471-2164-11-295.

Squair, Jordan W., Matthieu Gautier, Claudia Kathe, et al. 2021. "Confronting False Discoveries in Single-Cell Differential Expression." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-25960-2.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15: 1929–58. https://jmlr.org/papers/v15/srivastava14a.html.

Staden, R. 1979. "A Strategy of DNA Sequencing Employing Computer Programs." *Nucleic Acids Research* 6 (7): 2601–10. https://doi.org/10.1093/nar/6.7.2601.

Ståhl, Patrik L., Fredrik Salmén, Sanja Vickovic, et al. 2016. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics." *Science* 353 (6294): 78–82. https://doi.org/10.1126/science.aaf2403.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19 (suppl_2): ii215–25. https://doi.org/10.1093/bioinformatics/btg1080.

Stark, Rory, Marta Grzelak, and James Hadfield. 2019. "RNA Sequencing: The Teenage Years." *Nature Reviews Genetics* 20 (11): 631–56. https://doi.org/10.1038/s41576-019-0150-2.

Steel, Mike, and David Penny. 2000. "Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics." *Molecular Biology and Evolution* 17 (6): 839–50. https://doi.org/10.1093/oxfordjournals.molbev.a026364.

Steentoft, Catharina, Sergey Y Vakhrushev, Hiren J Joshi, et al. 2013. "Precision Mapping of the Human o-GalNAc Glycoproteome Through SimpleCell Technology." *The EMBO Journal* 32 (10): 1478–88. https://doi.org/10.1038/emboj.2013.79.

Steenwyk, Jacob L., Yuanning Li, Xiaofan Zhou, Xing-Xing Shen, and Antonis Rokas. 2023. "Incongruence in the Phylogenomics Era." *Nature Reviews Genetics* 24 (12): 834–50. https://doi.org/10.1038/s41576-023-00620-x.

Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. 2019. "HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation." *BMC Bioinformatics* 20 (1). https://doi.org/10.1186/s12859-019-3019-7.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28. https://doi.org/10.1038/nbt.3988.

Steinegger, Martin, and Johannes Söding. 2018. "Clustering Huge Protein Sequence Sets in Linear Time." *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-04964-5.

Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, et al. 2011. "Massive Genomic Rearrangement Acquired in a Single Catastrophic Event During Cancer Development." *Cell* 144 (1): 27–40. https://doi.org/10.1016/j.cell.2010.11.055.

Stevens, Jan A., Fabian Grünewald, P. A. Marco van Tilburg, et al. 2023. "Molecular Dynamics Simulation of an Entire Cell." *Frontiers in Chemistry* 11. https://doi.org/10.3389/fchem.2023.1106495.

Stiehler, Felix, Marvin Steinborn, Stephan Scholz, Daniela Dey, Andreas P M Weber, and Alisandra K Denton. 2020. "Helixer: Cross-Species Gene Annotation of Large Eukaryotic Genomes Using Deep Learning." *Bioinformatics* 36 (22-23): 5291–98. https://doi.org/10.1093/bioinformatics/btaa1044.

Stirling, David R., Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. 2021. "CellProfiler 4: Improvements in Speed, Utility and Usability." *BMC Bioinformatics* 22 (1). https://doi.org/10.1186/s12859-021-04344-9.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, et al. 2017. "Simultaneous Epitope and Transcriptome Measurement in Single Cells." *Nature Methods* 14 (9): 865–68. https://doi.org/10.1038/nmeth.4380.

Storer, Jessica, Robert Hubley, Jeb Rosen, Travis J. Wheeler, and Arian F. Smit. 2021. "The Dfam Community Resource of Transposable Element Families, Sequence Models, and Genome Annotations." *Mobile DNA* 12 (1). https://doi.org/10.1186/s13100-020-00230-y.

Storey, John D., and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences* 100 (16): 9440–45. https://doi.org/10.1073/pnas.1530509100.

Stormo, Gary D. 2000. "DNA Binding Sites: Representation and Discovery." *Bioinformatics* 16 (1): 16–23. https://doi.org/10.1093/bioinformatics/16.1.16.

Stormo, Gary D., Thomas D. Schneider, Larry Gold, and Andrzej Ehrenfeucht. 1982. "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in e. Coli." *Nucleic Acids Research* 10 (9): 2997–3011. https://doi.org/10.1093/nar/10.9.2997.

Storz, Gisela, Yuri I. Wolf, and Kumaran S. Ramamurthi. 2014. "Small Proteins Can No Longer Be Ignored." *Annual Review of Biochemistry* 83 (1): 753–77. https://doi.org/10.1146/annurev-biochem-070611-102400.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239): 719–24. https://doi.org/10.1038/nature07943.

Stringer, Carsen, Tim Wang, Michalis Michaelos, and Marius Pachitariu. 2020. "Cellpose: A Generalist Algorithm for Cellular Segmentation." *Nature Methods* 18 (1): 100–106. https://doi.org/10.1038/s41592-020-01018-x.

Strokach, Alexey, and Philip M. Kim. 2022. "Deep Generative Modeling for Protein Design." *Current Opinion in Structural Biology* 72: 226–36. https://doi.org/10.1016/j.sbi.2021.11.008.

Stryer, L, and R P Haugland. 1967. "Energy Transfer: A Spectroscopic Ruler." *Proceedings of the National Academy of Sciences* 58 (2): 719–26. https://doi.org/10.1073/pnas.58.2.719.

Studier, James A., and Karl J. Keppler. 1988. "A Note on the Neighbor-Joining Algorithm of Saitou and Nei." *Molecular Biology and Evolution* 5 (6): 729–31. https://doi.org/10.1093/oxfordjournals.molbev.a040527.

Su, Andrew I., Tim Wiltshire, Serge Batalov, et al. 2004. "A Gene Atlas of the Mouse and Human Protein-Encoding Transcriptomes." *Proceedings of the National Academy of Sciences* 101 (16): 6062–67. https://doi.org/10.1073/pnas.0400782101.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Subramanian, Indhupriya, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. 2020. "Multi-Omics Data Integration, Interpretation, and Its Application." *Bioinformatics and Biology Insights* 14: 117793221989905. https://doi.org/10.1177/1177932219899051.

Suchard, Marc A, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. 2018. "Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10." *Virus Evolution* 4 (1). https://doi.org/10.1093/ve/vey016.

Sugita, Yuji, and Yuko Okamoto. 1999. "Replica-Exchange Molecular Dynamics Method for Protein Folding." *Chemical Physics Letters* 314 (1-2): 141–51. https://doi.org/10.1016/s0009-2614(99)01123-9.

Sullivan, Patrick F., Jennifer R. S. Meadows, Steven Gazal, et al. 2023. "Leveraging Base-Pair Mammalian Constraint to Understand Genetic Variation and Human Disease." *Science* 380 (6643). https://doi.org/10.1126/science.abn2937.

Sun, Benjamin B., Joshua Chiou, Matthew Traylor, et al. 2023. "Plasma Proteomic Associations with Genetics and Health in the UK Biobank." *Nature* 622 (7982): 329–38. https://doi.org/10.1038/s41586-023-06592-6.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, et al. 2015. "Structure and Function of the Global Ocean Microbiome." *Science* 348 (6237). https://doi.org/10.1126/science.1261359.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." *PLoS ONE* 6 (7): e21800. https://doi.org/10.1371/journal.pone.0021800.

Sutherland, John D. 2015. "The Origin of Life—Out of the Blue." *Angewandte Chemie International Edition* 55 (1): 104–21. https://doi.org/10.1002/anie.201506585.

Suzek, Baris E., Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. and Wu. 2014. "UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches." *Bioinformatics* 31 (6): 926–32. https://doi.org/10.1093/bioinformatics/btu739.

Svensson, Valentine. 2020. "Droplet scRNA-Seq Is Not Zero-Inflated." *Nature Biotechnology* 38 (2): 147–50. https://doi.org/10.1038/s41587-019-0379-5.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A Teichmann. 2018. "Exponential Scaling of Single-Cell RNA-Seq in the Past Decade." *Nature Protocols* 13 (4): 599–604. https://doi.org/10.1038/nprot.2017.149.

Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, et al. 2022. "The STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest." *Nucleic Acids Research* 51 (D1): D638–46. https://doi.org/10.1093/nar/gkac1000.

Szklarczyk, Damian, Katerina Nastou, Mikaela Koutrouli, et al. 2024. "The STRING Database in 2025: Protein Networks with Directionality of Regulation." *Nucleic Acids Research* 53 (D1): D730–37. https://doi.org/10.1093/nar/gkae1113.

Szöllősi, Gergely J., Eric Tannier, Vincent Daubin, and Bastien Boussau. 2014. "The Inference of Gene Trees with Species Trees." *Systematic Biology* 64 (1): e42–62. https://doi.org/10.1093/sysbio/syu048.

Tajima, F. 1993. "Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis." *Genetics* 135 (2): 599–607. https://doi.org/10.1093/genetics/135.2.599.

Talavera, Gerard, and Jose Castresana. 2007. "Improvement of Phylogenies After Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments." *Systematic Biology* 56 (4): 564–77. https://doi.org/10.1080/10635150701472164.

Talevich, Eric, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. 2016. "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing." *PLOS Computational Biology* 12 (4): e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

Tanaka, Koichi, Hiroaki Waki, Yutaka Ido, et al. 1988. "Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of-Flight Mass Spectrometry." *Rapid Communications in Mass Spectrometry* 2 (8): 151–53. https://doi.org/10.1002/rcm.1290020802.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, et al. 2009. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nature Methods* 6 (5): 377–82. https://doi.org/10.1038/nmeth.1315.

Tarassov, Kirill, Vincent Messier, Christian R. Landry, et al. 2008. "An in Vivo Map of the Yeast Protein Interactome." *Science* 320 (5882): 1465–70. https://doi.org/10.1126/science.1153878.

Tarca, Adi Laurentiu, Sorin Draghici, Purvesh Khatri, et al. 2008. "A Novel Signaling Pathway Impact Analysis." *Bioinformatics* 25 (1): 75–82. https://doi.org/10.1093/bioinformatics/btn577.

Tarca, Adi L, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. 2007. "Machine Learning and Its Applications to Biology." *PLoS Computational Biology* 3 (6): e116. https://doi.org/10.1371/journal.pcbi.0030116.

Tatusov, Roman L., Eugene V. Koonin, and David J. Lipman. 1997. "A Genomic Perspective on Protein Families." *Science* 278 (5338): 631–37. https://doi.org/10.1126/science.278.5338.631.

Tatusova, Tatiana, Michael DiCuccio, Azat Badretdin, et al. 2016. "NCBI Prokaryotic Genome Annotation Pipeline." *Nucleic Acids Research* 44 (14): 6614–24. https://doi.org/10.1093/nar/gkw569.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The Evolutionary Origin of Orphan Genes." *Nature Reviews Genetics* 12 (10): 692–702. https://doi.org/10.1038/nrg3053.

Taverna, Darin M, and Richard A Goldstein. 2002. "Why Are Proteins so Robust to Site Mutations?" *Journal of Molecular Biology* 315 (3): 479–84. https://doi.org/10.1006/jmbi.2001.5226.

Taylor, Chris F, Norman W Paton, Kathryn S Lilley, et al. 2007. "The Minimum Information about a Proteomics Experiment (MIAPE)." *Nature Biotechnology* 25 (8): 887–93. https://doi.org/10.1038/nbt1329.

Taylor, Dylan J., Jordan M. Eizenga, Qiuhui Li, et al. 2024. "Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References." *Annual Review of Genomics and Human Genetics* 25 (1): 77–104. https://doi.org/10.1146/annurev-genom-021623-081639.

Tegnér, Jesper, M. K. Stephen Yeung, Jeff Hasty, and James J. Collins. 2003. "Reverse Engineering Gene Networks: Integrating Genetic Perturbations with Dynamical Modeling." *Proceedings of the National Academy of Sciences* 100 (10): 5944–49. https://doi.org/10.1073/pnas.0933416100.

Tegunov, Dimitry, Liang Xue, Christian Dienemann, Patrick Cramer, and Julia Mahamid. 2021. "Multi-Particle Cryo-EM Refinement with m Visualizes Ribosome-Antibiotic Complex at 3.5 å in Cells." *Nature Methods* 18 (2): 186–93. https://doi.org/10.1038/s41592-020-01054-7.

Teresa-Trueba, Irene de, Sara K. Goetz, Alexander Mattausch, et al. 2023. "Convolutional Networks for Supervised Mining of Molecular Patterns Within Cellular Context." *Nature Methods* 20 (2): 284–94. https://doi.org/10.1038/s41592-022-01746-2.

Terwilliger, Thomas C., Dorothee Liebschner, Tristan I. Croll, et al. 2023. "AlphaFold Predictions Are Valuable Hypotheses and Accelerate but Do Not Replace Experimental Structure Determination." *Nature Methods* 21 (1): 110–16. https://doi.org/10.1038/s41592-023-02087-4.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, et al. 2005. "Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae : Implications for the Microbial 'Pan-Genome'."

Proceedings of the National Academy of Sciences 102 (39): 13950–55. https://doi.org/10.1073/pnas.0506758102.

Teufel, Felix, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, et al. 2022. "SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models." *Nature Biotechnology* 40 (7): 1023–25. https://doi.org/10.1038/s41587-021-01156-3.

Teufel, Felix, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. 2023. "GraphPart: Homology Partitioning for Biological Sequence Analysis." *NAR Genomics and Bioinformatics* 5 (4). https://doi.org/10.1093/nargab/lqad088.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

The Arabidopsis Genome Initiative. 2000. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana." *Nature* 408 (6814): 796–815. https://doi.org/10.1038/35048692.

The C. elegans Sequencing Consortium. 1998. "Genome Sequence of the Nematode c. Elegans: A Platform for Investigating Biology." *Science* 282 (5396): 2012–18. https://doi.org/10.1126/science.282.5396.2012.

The Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87. https://doi.org/10.1038/nature04072.

The Critical Assessment of Genome Interpretation Consortium. 2024. "CAGI, the Critical Assessment of Genome Interpretation, Establishes Progress and Prospects for Computational Genetic Variant Interpretation Methods." *Genome Biology* 25 (1). https://doi.org/10.1186/s13059-023-03113-6.

The ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

The ENCODE Project Consortium. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710. https://doi.org/10.1038/s41586-020-2493-4.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, et al. 2023. "The Gene Ontology Knowledgebase in 2023." *GENETICS* 224 (1). https://doi.org/10.1093/genetics/iyad031.

The GTEx Consortium. 2020. "The GTEx Consortium Atlas of Genetic Regulatory Effects Across Human Tissues." *Science* 369 (6509): 1318–30. https://doi.org/10.1126/science.aaz1776.

The Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.

The Tabula Muris Consortium. 2018. "Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris." *Nature* 562 (7727): 367–72. https://doi.org/10.1038/s41586-018-0590-4.

The Tabula Sapiens Consortium. 2022. "The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans." *Science* 376 (6594). https://doi.org/10.1126/science.abl4896.

Theodoris, Christina V., Ling Xiao, Anant Chopra, et al. 2023. "Transfer Learning Enables Predictions in

Network Biology." *Nature* 618 (7965): 616–24. https://doi.org/10.1038/s41586-023-06139-9.

Thiele, Ines, and Bernhard Ø Palsson. 2010. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction." *Nature Protocols* 5 (1): 93–121. https://doi.org/10.1038/nprot.2009.203.

Thomas, C. A. 1971. "THE GENETIC ORGANIZATION OF CHROMOSOMES." *Annual Review of Genetics* 5 (1): 237–56. https://doi.org/10.1146/annurev.ge.05.120171.001321.

Thomas, Gregg W. C., and Matthew W. Hahn. 2015. "Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study Using Echolocating Mammals." *Molecular Biology and Evolution* 32 (5): 1232–36. https://doi.org/10.1093/molbev/msv013.

Thomas, Paul D. 2016. "The Gene Ontology and the Meaning of Biological Function." In *The Gene Ontology Handbook*. Springer New York. https://doi.org/10.1007/978-1-4939-3743-1_2.

Thomas, Paul D., and Ken A. Dill. 1996. "Statistical Potentials Extracted from Protein Structures: How Accurate Are They?" *Journal of Molecular Biology* 257 (2): 457–69. https://doi.org/10.1006/jmbi.1996.0175.

Thomas, Paul D., David P. Hill, Huaiyu Mi, et al. 2019. "Gene Ontology Causal Activity Modeling (GO-CAM) Moves Beyond GO Annotations to Structured Descriptions of Biological Functions and Systems." *Nature Genetics* 51 (10): 1429–33. https://doi.org/10.1038/s41588-019-0500-1.

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, et al. 2003. "Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS." *Analytical Chemistry* 75 (8): 1895–904. https://doi.org/10.1021/ac0262560.

Thompson, J D, F Plewniak, and O Poch. 1999. "BAliBASE: A Benchmark Alignment Database for the Evaluation of Multiple Alignment Programs." *Bioinformatics* 15 (1): 87–88. https://doi.org/10.1093/bioinformatics/15.1.87.

Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. 1994. "CLUSTAL w: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." *Nucleic Acids Research* 22 (22): 4673–80. https://doi.org/10.1093/nar/22.22.4673.

Thornburg, Zane R., David M. Bianchi, Troy A. Brier, et al. 2022. "Fundamental Behaviors Emerge from Simulations of a Living Minimal Cell." *Cell* 185 (2): 345–360.e28. https://doi.org/10.1016/j.cell.2021.12.025.

Thornburg, Zane R., Andrew Maytin, Jiwoong Kwon, et al. 2026. "Bringing the Genetically Minimal Cell to Life on a Computer in 4D." *Cell* 189 (9): 2582–2597.e27. https://doi.org/10.1016/j.cell.2026.02.009.

Thorne, Jeffrey L., Hirohisa Kishino, and Joseph Felsenstein. 1991. "An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences." *Journal of Molecular Evolution* 33 (2): 114–24. https://doi.org/10.1007/bf02193625.

Thornton, Joseph W. 2004. "Resurrecting Ancient Genes: Experimental Analysis of Extinct Molecules." *Nature Reviews Genetics* 5 (5): 366–75. https://doi.org/10.1038/nrg1324.

Thorp, H. Holden. 2023. "ChatGPT Is Fun, but Not an Author." *Science* 379 (6630): 313–13. https:

//doi.org/10.1126/science.adg7879.

Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, et al. 2017. "A Subcellular Map of the Human Proteome." *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.

Thumuluri, Vineet, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. 2022. "DeepLoc 2.0: Multi-Label Subcellular Localization Prediction Using Protein Language Models." *Nucleic Acids Research* 50 (W1): W228–34. https://doi.org/10.1093/nar/gkac278.

Tian, Weidong, and Jeffrey Skolnick. 2003. "How Well Is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?" *Journal of Molecular Biology* 333 (4): 863–82. https://doi.org/10.1016/j.jmb.2003.08.057.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. 2002. "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression." *Proceedings of the National Academy of Sciences* 99 (10): 6567–72. https://doi.org/10.1073/pnas.082099299.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 63 (2): 411–23. https://doi.org/10.1111/1467-9868.00293.

Timmis, Jeremy N., Michael A. Ayliffe, Chun Y. Huang, and William Martin. 2004. "Endosymbiotic Gene Transfer: Organelle Genomes Forge Eukaryotic Chromosomes." *Nature Reviews Genetics* 5 (2): 123–35. https://doi.org/10.1038/nrg1271.

Timp, Winston, and Gregory Timp. 2020. "Beyond Mass Spectrometry, the Next Step in Proteomics." *Science Advances* 6 (2). https://doi.org/10.1126/sciadv.aax8978.

Ting, Lily, Ramin Rad, Steven P Gygi, and Wilhelm Haas. 2011. "MS3 Eliminates Ratio Distortion in Isobaric Multiplexed Quantitative Proteomics." *Nature Methods* 8 (11): 937–40. https://doi.org/10.1038/nmeth.1714.

Tirion, Monique M. 1996. "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." *Physical Review Letters* 77 (9): 1905–8. https://doi.org/10.1103/physrevlett.77.1905.

Tirosh, Itay, Benjamin Izar, Sanjay M. Prakadan, et al. 2016. "Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq." *Science* 352 (6282): 189–96. https://doi.org/10.1126/science.aad0501.

Toby, Timothy K., Luca Fornelli, and Neil L. Kelleher. 2016. "Progress in Top-down Proteomics and the Analysis of Proteoforms." *Annual Review of Analytical Chemistry* 9 (1): 499–519. https://doi.org/10.1146/annurev-anchem-071015-041550.

Togashi, Yuichi, and Holger Flechsig. 2018. "Coarse-Grained Protein Dynamics Studies Using Elastic Network Models." *International Journal of Molecular Sciences* 19 (12): 3899. https://doi.org/10.3390/ijms19123899.

Toro, Noemi del, Anjali Shrivastava, Eliot Ragueneau, et al. 2021. "The IntAct Database: Efficient Access to Fine-Grained Molecular Interaction Data." *Nucleic Acids Research* 50 (D1): D648–53. https://doi.org/10.1093/nar/gkab1006.

Touchon, Marie, Claire Hoede, Olivier Tenaillon, et al. 2009. "Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths." *PLoS Genetics* 5 (1): e1000344. https://doi.org/10.1371/journal.pgen.1000344.

Touw, W. G., J. R. Bayjanov, L. Overmars, et al. 2012. "Data Mining in the Life Sciences with Random Forest: A Walk in the Park or Lost in the Jungle?" *Briefings in Bioinformatics* 14 (3): 315–26. https://doi.org/10.1093/bib/bbs034.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-41695-z.

Tran, John C., Leonid Zamdborg, Dorothy R. Ahlf, et al. 2011. "Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics." *Nature* 480 (7376): 254–58. https://doi.org/10.1038/nature10575.

Tran, Ngoc Hieu, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. 2017. "De Novo Peptide Sequencing by Deep Learning." *Proceedings of the National Academy of Sciences* 114 (31): 8247–52. https://doi.org/10.1073/pnas.1705691114.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, et al. 2014. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells." *Nature Biotechnology* 32 (4): 381–86. https://doi.org/10.1038/nbt.2859.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics* 25 (9): 1105–11. https://doi.org/10.1093/bioinformatics/btp120.

Trapnell, Cole, Brian A Williams, Geo Pertea, et al. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation." *Nature Biotechnology* 28 (5): 511–15. https://doi.org/10.1038/nbt.1621.

Treangen, Todd J., and Eduardo P. C. Rocha. 2011. "Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes." *PLoS Genetics* 7 (1): e1001284. https://doi.org/10.1371/journal.pgen.1001284.

Tress, Michael L., Federico Abascal, and Alfonso Valencia. 2017. "Alternative Splicing May Not Be the Key to Proteome Complexity." *Trends in Biochemical Sciences* 42 (2): 98–110. https://doi.org/10.1016/j.tibs.2016.08.008.

Trott, Oleg, and Arthur J. Olson. 2009. "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading." *Journal of Computational Chemistry* 31 (2): 455–61. https://doi.org/10.1002/jcc.21334.

Tsherniak, Aviad, Francisca Vazquez, Phil G. Montgomery, et al. 2017. "Defining a Cancer Dependency Map." *Cell* 170 (3): 564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010.

Tsoka, Sophia, and Christos A. Ouzounis. 2000. "Prediction of Protein Interactions: Metabolic Enzymes Are Frequently Involved in Gene Fusion." *Nature Genetics* 26 (2): 141–42. https://doi.org/10.1038/79

847.

Tsuchiya, Yuko, Yu Yamamori, and Kentaro Tomii. 2022. "Protein–Protein Interaction Prediction Methods: From Docking-Based to AI-Based Approaches." *Biophysical Reviews* 14 (6): 1341–48. https://doi.org/10.1007/s12551-022-01032-7.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome." *Nature* 596 (7873): 590–96. https://doi.org/10.1038/s41586-021-03828-1.

Turner, Jack, Sanja Abbott, Neli Fonseca, et al. 2023. "EMDB—the Electron Microscopy Data Bank." *Nucleic Acids Research* 52 (D1): D456–65. https://doi.org/10.1093/nar/gkad1019.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, et al. 2004. "Community Structure and Metabolism Through Reconstruction of Microbial Genomes from the Environment." *Nature* 428 (6978): 37–43. https://doi.org/10.1038/nature02340.

Tyson, John J, Katherine C Chen, and Bela Novak. 2003. "Sniffers, Buzzers, Toggles and Blinkers: Dynamics of Regulatory and Signaling Pathways in the Cell." *Current Opinion in Cell Biology* 15 (2): 221–31. https://doi.org/10.1016/s0955-0674(03)00017-6.

Uetz, Peter, Loic Giot, Gerard Cagney, et al. 2000. "A Comprehensive Analysis of Protein–Protein Interactions in Saccharomyces Cerevisiae." *Nature* 403 (6770): 623–27. https://doi.org/10.1038/35001009.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, et al. 2021. "Genome-Wide Association Studies." *Nature Reviews Methods Primers* 1 (1). https://doi.org/10.1038/s43586-021-00056-9.

Uhlen, Mathias, Anita Bandrowski, Steven Carr, et al. 2016. "A Proposal for Validation of Antibodies." *Nature Methods* 13 (10): 823–27. https://doi.org/10.1038/nmeth.3995.

Uhlen, Mathias, Cheng Zhang, Sunjae Lee, et al. 2017. "A Pathology Atlas of the Human Cancer Transcriptome." *Science* 357 (6352). https://doi.org/10.1126/science.aan2507.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, et al. 2015. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220). https://doi.org/10.1126/science.1260419.

Vakirlis, Nikolaos, Anne-Ruxandra Carvunis, and Aoife McLysaght. 2020. "Synteny-Based Analyses Indicate That Sequence Divergence Is Not the Main Source of Orphan Genes." *eLife* 9. https://doi.org/10.7554/elife.53500.

Valentini, Giorgio, Dario Malchiodi, Jessica Gliozzo, et al. 2023. "The Promises of Large Language Models for Protein Design and Modeling." *Frontiers in Bioinformatics* 3. https://doi.org/10.3389/fbinf.2023.1304099.

Van Buren, Samantha L, C Titus Brown, Tomasz Szmatoła, and Carrie J Finno. 2025. "A Comparative Review of Short Genetic Variant Databases Across Humans and Animal Species." *Briefings in Bioinformatics* 26 (4). https://doi.org/10.1093/bib/bbaf356.

van Kempen, Michel, Stephanie S. Kim, Charlotte Tumescheit, et al. 2023. "Fast and Accurate Protein Structure Search with Foldseek." *Nature Biotechnology* 42 (2): 243–46. https://doi.org/10.1038/s41587-

023-01773-0.

Van Nostrand, Eric L, Gabriel A Pratt, Alexander A Shishkin, et al. 2016. "Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP)." *Nature Methods* 13 (6): 508–14. https://doi.org/10.1038/nmeth.3810.

Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. "De Novo Gene Birth." *PLOS Genetics* 15 (5): e1008160. https://doi.org/10.1371/journal.pgen.1008160.

Vanunu, Oron, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. 2010. "Associating Genes and Protein Complexes with Disease via Network Propagation." *PLoS Computational Biology* 6 (1): e1000641. https://doi.org/10.1371/journal.pcbi.1000641.

Varadi, Mihaly, Stephen Anyango, Sri Devan Appasamy, et al. 2022. "PDBe and PDBe-KB: Providing High-Quality, up-to-Date and Integrated Resources of Macromolecular Structures to Support Basic and Applied Research and Education." *Protein Science* 31 (10). https://doi.org/10.1002/pro.4439.

Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, et al. 2021. "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models." *Nucleic Acids Research* 50 (D1): D439–44. https://doi.org/10.1093/nar/gkab1061.

Varadi, Mihaly, Damian Bertoni, Paulyna Magana, et al. 2023. "AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences." *Nucleic Acids Research* 52 (D1): D368–75. https://doi.org/10.1093/nar/gkad1011.

Varma, Sudhir, and Richard Simon. 2006. "Bias in Error Estimation When Using Cross-Validation for Model Selection." *BMC Bioinformatics* 7 (1). https://doi.org/10.1186/1471-2105-7-91.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. https://arxiv.org/abs/1706.03762.

Veer, Laura J. van 't, Hongyue Dai, Marc J. van de Vijver, et al. 2002. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature* 415 (6871): 530–36. https://doi.org/10.1038/415530a.

Velculescu, Victor E., Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. 1995. "Serial Analysis of Gene Expression." *Science* 270 (5235): 484–87. https://doi.org/10.1126/science.270.5235.484.

Veltman, Joris A., and Han G. Brunner. 2012. "De Novo Mutations in Human Genetic Disease." *Nature Reviews Genetics* 13 (8): 565–75. https://doi.org/10.1038/nrg3241.

Vendramin, Roberto, Kevin Litchfield, and Charles Swanton. 2021. "Cancer Evolution: Darwin and Beyond." *The EMBO Journal* 40 (18). https://doi.org/10.15252/embj.2021108389.

Venet, David, Jacques E. Dumont, and Vincent Detours. 2011. "Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7 (10): e1002240. https://doi.org/10.1371/journal.pcbi.1002240.

Vénien-Bryan, Catherine, Zhuolun Li, Laurent Vuillard, and Jean Albert Boutin. 2017. "Cryo-Electron Microscopy and x-Ray Crystallography: Complementary Approaches to Structural Biology and Drug Discovery." *Acta Crystallographica Section F Structural Biology Communications* 73 (4): 174–83. https://doi.org/10.1107/s2053230x17003740.

Venkatesan, Kavitha, Jean-François Rual, Alexei Vazquez, et al. 2008. "An Empirical Framework for Binary Interactome Mapping." *Nature Methods* 6 (1): 83–90. https://doi.org/10.1038/nmeth.1280.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Venter, J. Craig, Karin Remington, John F. Heidelberg, et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304 (5667): 66–74. https://doi.org/10.1126/science.1093857.

Viana, Matheus P., Jianxu Chen, Theo A. Knijnenburg, et al. 2023. "Integrated Intracellular Organization and Its Variations in Human iPS Cells." *Nature* 613 (7943): 345–54. https://doi.org/10.1038/s41586-022-05563-7.

Vilella, Albert J., Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. 2008. "EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates." *Genome Research* 19 (2): 327–35. https://doi.org/10.1101/gr.073585.107.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, et al. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *The American Journal of Human Genetics* 101 (1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Viterbi, A. 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory* 13 (2): 260–69. https://doi.org/10.1109/tit.1967.1054010.

Vizcaíno, Juan A, Eric W Deutsch, Rui Wang, et al. 2014. "ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination." *Nature Biotechnology* 32 (3): 223–26. https://doi.org/10.1038/nbt.2839.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58. https://doi.org/10.1126/science.1235122.

Vorontsov, Eugene, Alican Bozkurt, Adam Casson, et al. 2024. "A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection." *Nature Medicine* 30 (10): 2924–35. https://doi.org/10.1038/s41591-024-03141-0.

Vu, Thi Thuy Duong, and Jaehee Jung. 2021. "Protein Function Prediction with Gene Ontology: From Traditional to Deep Learning Models." *PeerJ* 9: e12019. https://doi.org/10.7717/peerj.12019.

Vuong, Paton, Michael J. Wise, Andrew S. Whiteley, and Parwinder Kaur. 2022. "Ten Simple Rules for Investigating (Meta)genomic Data from Environmental Ecosystems." *PLOS Computational Biology* 18 (12): e1010675. https://doi.org/10.1371/journal.pcbi.1010675.

Wächtershäuser, G. 1988. "Before Enzymes and Templates: Theory of Surface Metabolism." *Microbiological Reviews* 52 (4): 452–84. https://doi.org/10.1128/mr.52.4.452-484.1988.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. "Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples." *Theory in Biosciences* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Walsh, Ian, Dmytro Fishman, Dario Garcia-Gasulla, et al. 2021. "DOME: Recommendations for Supervised Machine Learning Validation in Biology." *Nature Methods* 18 (10): 1122–27. https://doi.org/10.1038/s41592-021-01205-4.

Waman, Vaishali P., Nicola Bordin, Rachel Alcraft, et al. 2024. "CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds." *Journal of Molecular Biology* 436 (17): 168551. https://doi.org/10.1016/j.jmb.2024.168551.

Wan, Cuihong, Blake Borgeson, Sadhna Phanse, et al. 2015. "Panorama of Ancient Metazoan Macromolecular Complexes." *Nature* 525 (7569): 339–44. https://doi.org/10.1038/nature14877.

Wang, Bo, Aziz M Mezlini, Feyyaz Demir, et al. 2014. "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale." *Nature Methods* 11 (3): 333–37. https://doi.org/10.1038/nmeth.2810.

Wang, Duolin, Dongpeng Liu, Jiakang Yuchi, et al. 2020. "MusiteDeep: A Deep-Learning Based Webserver for Protein Post-Translational Modification Site Prediction and Visualization." *Nucleic Acids Research* 48 (W1): W140–46. https://doi.org/10.1093/nar/gkaa275.

Wang, Eric T., Rickard Sandberg, Shujun Luo, et al. 2008. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456 (7221): 470–76. https://doi.org/10.1038/nature07509.

Wang, Jue, Sidney Lisanza, David Juergens, et al. 2022. "Scaffolding Protein Functional Sites Using Deep Learning." *Science* 377 (6604): 387–94. https://doi.org/10.1126/science.abn2100.

Wang, Lusheng, and Tao Jiang. 1994. "On the Complexity of Multiple Sequence Alignment." *Journal of Computational Biology* 1 (4): 337–48. https://doi.org/10.1089/cmb.1994.1.337.

Wang, Sheng, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. 2017. "Accurate de Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model." *PLOS Computational Biology* 13 (1): e1005324. https://doi.org/10.1371/journal.pcbi.1005324.

Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." *Science* 343 (6166): 80–84. https://doi.org/10.1126/science.1246981.

Wang, Yan, Hang Zhang, Haolin Zhong, and Zhidong Xue. 2021. "Protein Domain Identification Methods and Online Resources." *Computational and Structural Biotechnology Journal* 19: 1145–53. https://doi.org/10.1016/j.csbj.2021.01.041.

Wang, Yen-Wen, Jaqueline Hess, Jason C Slot, and Anne Pringle. 2020. "De Novo Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis inAmanita." *Genome Biology and Evolution* 12 (11): 2168–82. https://doi.org/10.1093/gbe/evaa193.

Wang, Y., H. Tang, J. D. DeBarry, et al. 2012. "MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity." *Nucleic Acids Research* 40 (7): e49–49. https://doi.org/10.1093/nar/gkr1293.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Wang, Zichen, Alexander Lachmann, and Avi Ma'ayan. 2018. "Mining Data and Metadata from the Gene Expression Omnibus." *Biophysical Reviews* 11 (1): 103–10. https://doi.org/10.1007/s12551-018-0490-8.

Wang, Zicong, Peng-Kai Liu, and Lingjun Li. 2024. "A Tutorial Review of Labeling Methods in Mass Spectrometry-Based Quantitative Proteomics." *ACS Measurement Science Au* 4 (4): 315–37. https://doi.org/10.1021/acsmeasuresciau.4c00007.

Warren, Gregory L., C. Webster Andrews, Anna-Maria Capelli, et al. 2005. "A Critical Assessment of Docking Programs and Scoring Functions." *Journal of Medicinal Chemistry* 49 (20): 5912–31. https://doi.org/10.1021/jm050362n.

Warshel, A., and M. Levitt. 1976. "Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme." *Journal of Molecular Biology* 103 (2): 227–49. https://doi.org/10.1016/0022-2836(76)90311-9.

Washburn, Michael P., Dirk Wolters, and John R. Yates. 2001. "Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology." *Nature Biotechnology* 19 (3): 242–47. https://doi.org/10.1038/85686.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA Statement on p -Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Watson, Joseph L., David Juergens, Nathaniel R. Bennett, et al. 2023. "De Novo Design of Protein Structure and Function with RFdiffusion." *Nature* 620 (7976): 1089–100. https://doi.org/10.1038/s41586-023-06415-8.

Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393 (6684): 440–42. https://doi.org/10.1038/30918.

Wayment-Steele, Hannah K., Adedolapo Ojoawo, Renee Otten, et al. 2023. "Predicting Multiple Conformations via Sequence Clustering and AlphaFold2." *Nature* 625 (7996): 832–39. https://doi.org/10.1038/s41586-023-06832-9.

Weber, James L., and Eugene W. Myers. 1997. "Human Whole-Genome Shotgun Sequencing." *Genome Research* 7 (5): 401–9. https://doi.org/10.1101/gr.7.5.401.

Weber, Lukas M., Wouter Saelens, Robrecht Cannoodt, et al. 2019. "Essential Guidelines for Computational Method Benchmarking." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1738-8.

Weigt, Martin, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. 2009. "Identification of Direct Residue Contacts in Protein–Protein Interaction by Message Passing." *Proceedings of the National Academy of Sciences* 106 (1): 67–72. https://doi.org/10.1073/pnas.0805923106.

Weinstein, John N, Eric A Collisson, Gordon B Mills, et al. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* 45 (10): 1113–20. https://doi.org/10.1038/ng.2764.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2020. "Many, but Not All, Lineage-

Specific Genes Can Be Explained by Homology Detection Failure." *PLOS Biology* 18 (11): e3000862. https://doi.org/10.1371/journal.pbio.3000862.

Weiss, Madeline C., Filipa L. Sousa, Natalia Mrnjavac, et al. 2016. "The Physiology and Habitat of the Last Universal Common Ancestor." *Nature Microbiology* 1 (9). https://doi.org/10.1038/nmicrobiol.2016.116.

Weissenow, Konstantin, and Burkhard Rost. 2025. "Are Protein Language Models the New Universal Key?" *Current Opinion in Structural Biology* 91: 102997. https://doi.org/10.1016/j.sbi.2025.102997.

Wells, Jonathan N., and Cédric Feschotte. 2020. "A Field Guide to Eukaryotic Transposable Elements." *Annual Review of Genetics* 54 (1): 539–61. https://doi.org/10.1146/annurev-genet-040620-022145.

Wen, Bo, Jack Freestone, Michael Riffle, Michael J. MacCoss, William S. Noble, and Uri Keich. 2025. "Assessment of False Discovery Rate Control in Tandem Mass Spectrometry Analysis Using Entrapment." *Nature Methods* 22 (7): 1454–63. https://doi.org/10.1038/s41592-025-02719-x.

Wen, Bo, Wen-Feng Zeng, Yuxing Liao, et al. 2020. "Deep Learning in Proteomics." *PROTEOMICS* 20 (21-22). https://doi.org/10.1002/pmic.201900335.

Wenger, Aaron M., Paul Peluso, William J. Rowell, et al. 2019. "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature Biotechnology* 37 (10): 1155–62. https://doi.org/10.1038/s41587-019-0217-9.

Wetlaufer, Donald B. 1973. "Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins." *Proceedings of the National Academy of Sciences* 70 (3): 697–701. https://doi.org/10.1073/pnas.70.3.697.

Wheeler, Travis J., and Sean R. Eddy. 2013. "Nhmmer: DNA Homology Search with Profile HMMs." *Bioinformatics* 29 (19): 2487–89. https://doi.org/10.1093/bioinformatics/btt403.

Whelan, Simon, and Nick Goldman. 2001. "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach." *Molecular Biology and Evolution* 18 (5): 691–99. https://doi.org/10.1093/oxfordjournals.molbev.a003851.

Whisstock, James C., and Arthur M. Lesk. 2003. "Prediction of Protein Function from Protein Sequence and Structure." *Quarterly Reviews of Biophysics* 36 (3): 307–40. https://doi.org/10.1017/s0033583503003901.

Whitley, Paul, Brayan Grau, James C. Gumbart, Luis Martínez-Gil, and Ismael Mingarro. 2021. "Folding and Insertion of Transmembrane Helices at the ER." *International Journal of Molecular Sciences* 22 (23): 12778. https://doi.org/10.3390/ijms222312778.

Wickner, Reed B. 1994. "[URE3] as an Altered URE2 Protein: Evidence for a Prion Analog in Saccharomyces Cerevisiae." *Science* 264 (5158): 566–69. https://doi.org/10.1126/science.7909170.

Wijesooriya, Kaumadi, Sameer A. Jadaan, Kaushalya L. Perera, Tanuveer Kaur, and Mark Ziemann. 2022. "Urgent Need for Consistent Standards in Functional Enrichment Analysis." *PLOS Computational Biology* 18 (3): e1009935. https://doi.org/10.1371/journal.pcbi.1009935.

Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, et al. 2014. "Mass-Spectrometry-Based Draft of the Human Proteome." *Nature* 509 (7502): 582–87. https://doi.org/10.1038/nature13319.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1). https://doi.org/10.1038/sdata.2016.18.

Will, Sebastian, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. 2007. "Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering." *PLoS Computational Biology* 3 (4): e65. https://doi.org/10.1371/journal.pcbi.0030065.

Williams, Cameron G., Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. 2022. "An Introduction to Spatial Transcriptomics for Biomedical Research." *Genome Medicine* 14 (1). https://doi.org/10.1186/s13073-022-01075-1.

Williams, Eleanor, Josh Moore, Simon W Li, et al. 2017. "Image Data Resource: A Bioimage Data Integration and Publication Platform." *Nature Methods* 14 (8): 775–81. https://doi.org/10.1038/nmeth.4326.

Williams, Paul D, David D Pollock, Benjamin P Blackburne, and Richard A Goldstein. 2006. "Assessing the Accuracy of Ancestral Protein Reconstruction Methods." *PLoS Computational Biology* 2 (6): e69. https://doi.org/10.1371/journal.pcbi.0020069.

Williams, Tom A., Cymon J. Cox, Peter G. Foster, Gergely J. Szöllősi, and T. Martin Embley. 2019. "Phylogenomics Provides Robust Support for a Two-Domains Tree of Life." *Nature Ecology & Evolution* 4 (1): 138–47. https://doi.org/10.1038/s41559-019-1040-x.

Williamson, Michael P., Timothy F. Havel, and Kurt Wüthrich. 1985. "Solution Conformation of Proteinase Inhibitor IIA from Bull Seminal Plasma by 1H Nuclear Magnetic Resonance and Distance Geometry." *Journal of Molecular Biology* 182 (2): 295–315. https://doi.org/10.1016/0022-2836(85)90347-x.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. "Good Enough Practices in Scientific Computing." *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

Wimalaratne, Sarala M., Nick Juty, John Kunze, et al. 2018. "Uniform Resolution of Compact Identifiers for Biomedical Data." *Scientific Data* 5 (1). https://doi.org/10.1038/sdata.2018.29.

Wimley, William C. 2003. "The Versatile β-Barrel Membrane Protein." *Current Opinion in Structural Biology* 13 (4): 404–11. https://doi.org/10.1016/s0959-440x(03)00099-x.

Winnifrith, Adam, Carlos Outeiral, and Brian L. Hie. 2024. "Generative Artificial Intelligence for de Novo Protein Design." *Current Opinion in Structural Biology* 86: 102794. https://doi.org/10.1016/j.sbi.2024.102794.

Winzeler, Elizabeth A., Daniel D. Shoemaker, Anna Astromoff, et al. 1999. "Functional Characterization of the s. Cerevisiae Genome by Gene Deletion and Parallel Analysis." *Science* 285 (5429): 901–6. https://doi.org/10.1126/science.285.5429.901.

Woese, C R, O Kandler, and M L Wheelis. 1990. "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya." *Proceedings of the National Academy of Sciences* 87 (12): 4576–79. https://doi.org/10.1073/pnas.87.12.4576.

Woese, Carl R., and George E. Fox. 1977. "Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms." *Proceedings of the National Academy of Sciences* 74 (11): 5088–90. https://doi.org/10.1073/pnas.74.11.5088.

Wohlwend, Jeremy, Gabriele Corso, Saro Passaro, et al. 2024. "Boltz-1 Democratizing Biomolecular Interaction Modeling." *bioRxiv (Preprint)*, ahead of print. https://doi.org/10.1101/2024.11.19.624167.

Wolf, Aaron B., and Joshua M. Akey. 2018. "Outstanding Questions in the Study of Archaic Hominin Admixture." *PLOS Genetics* 14 (5): e1007349. https://doi.org/10.1371/journal.pgen.1007349.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1). https://doi.org/10.1186/s13059-017-1382-0.

Wolfe, Kenneth H., and Denis C. Shields. 1997. "Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome." *Nature* 387 (6634): 708–13. https://doi.org/10.1038/42711.

Wong, Karen M., Marc A. Suchard, and John P. Huelsenbeck. 2008. "Alignment Uncertainty and Genomic Analysis." *Science* 319 (5862): 473–76. https://doi.org/10.1126/science.1151532.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1891-0.

Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3). https://doi.org/10.1186/gb-2014-15-3-r46.

Wooley, John C., Adam Godzik, and Iddo Friedberg. 2010. "A Primer on Metagenomics." *PLoS Computational Biology* 6 (2): e1000667. https://doi.org/10.1371/journal.pcbi.1000667.

Wootton, John C., and Scott Federhen. 1993. "Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases." *Computers & Chemistry* 17 (2): 149–63. https://doi.org/10.1016/0097-8485(93)85006-x.

Xiao, Hanyu, Yijin Zou, Jieqiong Wang, and Shibiao Wan. 2024. "A Review for Artificial Intelligence Based Protein Subcellular Localization." *Biomolecules* 14 (4): 409. https://doi.org/10.3390/biom14040409.

Xie, Yubin, Yueyuan Zheng, Hongyu Li, et al. 2016. "GPS-Lipid: A Robust Tool for the Prediction of Multiple Lipid Modification Sites." *Scientific Reports* 6 (1). https://doi.org/10.1038/srep28249.

Xu, Hanwen, Naoto Usuyama, Jaspreet Bagga, et al. 2024. "A Whole-Slide Foundation Model for Digital Pathology from Real-World Data." *Nature* 630 (8015): 181–88. https://doi.org/10.1038/s41586-024-07441-w.

Yanai, Itai, and Charles DeLisi. 2002. "The Society of Genes: Networks of Functional Links Between Genes from Comparative Genomics." *Genome Biology* 3 (11). https://doi.org/10.1186/gb-2002-3-11-research0064.

Yancopoulos, S., O. Attie, and R. Friedberg. 2005. "Efficient Sorting of Genomic Permutations by Translocation, Inversion and Block Interchange." *Bioinformatics* 21 (16): 3340–46. https://doi.org/10.1093/bioinformatics/bti535.

Yandell, Mark, and Daniel Ence. 2012. "A Beginner's Guide to Eukaryotic Genome Annotation." *Nature Reviews Genetics* 13 (5): 329–42. https://doi.org/10.1038/nrg3174.

Yang, Jian Le, Qing Li, Sen Zhao, et al. 2026. "Molecular Diagnostic Yield of Exome Sequencing and Genome Sequencing in Critical Ill Neonates and Infants: A Systematic Review and Meta-Analysis." *Genetics in Medicine* 28 (8): 102627. https://doi.org/10.1016/j.gim.2026.102627.

Yang, Jian, Beben Benyamin, Brian P McEvoy, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–69. https://doi.org/10.1038/ng.608.

Yang, Z. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Yang, Ziheng. 1994. "Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods." *Journal of Molecular Evolution* 39 (3): 306–14. https://doi.org/10.1007/bf00160154.

Yang, Ziheng, and Joseph P. Bielawski. 2000. "Statistical Methods for Detecting Molecular Adaptation." *Trends in Ecology & Evolution* 15 (12): 496–503. https://doi.org/10.1016/s0169-5347(00)01994-7.

Yang, Ziheng, and Bruce Rannala. 2012. "Molecular Phylogenetics: Principles and Practice." *Nature Reviews Genetics* 13 (5): 303–14. https://doi.org/10.1038/nrg3186.

Yang, Z, S Kumar, and M Nei. 1995. "A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences." *Genetics* 141 (4): 1641–50. https://doi.org/10.1093/genetics/141.4.1641.

Yilmaz, Melih, William E. Fondrie, Wout Bittremieux, et al. 2024. "Sequence-to-Sequence Translation from Mass Spectra to Peptides with a Transformer Model." *Nature Communications* 15 (1). https://doi.org/10.1038/s41467-024-49731-x.

Yoo, DongAhn, Arang Rhie, Prajna Hebbar, et al. 2025. "Complete Sequencing of Ape Genomes." *Nature* 641 (8062): 401–18. https://doi.org/10.1038/s41586-025-08816-3.

Yoon, Byung-Jun. 2009. "Hidden Markov Models and Their Applications in Biological Sequence Analysis." *Current Genomics* 10 (6): 402–15. https://doi.org/10.2174/138920209789177575.

You, Yue, Yuting Fu, Lanxiang Li, et al. 2024. "Systematic Comparison of Sequencing-Based Spatial Transcriptomic Methods." *Nature Methods* 21 (9): 1743–54. https://doi.org/10.1038/s41592-024-02325-3.

Young, Lindsey N., and Elizabeth Villa. 2023. "Bringing Structure to Cell Biology with Cryo-Electron Tomography." *Annual Review of Biophysics* 52 (1): 573–95. https://doi.org/10.1146/annurev-biophys-111622-091327.

Youssef, Noor, Aidan Budd, and Joseph P. Bielawski. 2019. "Introduction to Genome Biology and Diversity." In *Evolutionary Genomics*. Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_1.

Yu, Haiyuan, Pascal Braun, Muhammed A. Yıldırım, et al. 2008. "High-Quality Binary Protein Interaction Map of the Yeast Interactome Network." *Science* 322 (5898): 104–10. https://doi.org/10.1126/science.1158684.

Yu, Isseki, Takaharu Mori, Tadashi Ando, et al. 2016. "Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm." *eLife* 5. https://doi.org/10.7554/elife.19274.

Yu, Tianhao, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. 2023. "Enzyme Function Prediction Using Contrastive Learning." *Science* 379 (6639): 1358–63. https://doi.org/10.1126/science.adf2465.

Yuan, Rongqing, Jing Zhang, Andriy Kryshtafovych, et al. 2025. "CASP16 Protein Monomer Structure Prediction Assessment." *Proteins: Structure, Function, and Bioinformatics* 94 (1): 86–105. https://doi.org/10.1002/prot.70031.

Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, et al. 2017. "Asgard Archaea Illuminate the Origin of Eukaryotic Cellular Complexity." *Nature* 541 (7637): 353–58. https://doi.org/10.1038/nature21031.

Zemla, A. 2003. "LGA: A Method for Finding 3D Similarities in Protein Structures." *Nucleic Acids Research* 31 (13): 3370–74. https://doi.org/10.1093/nar/gkg571.

Zemla, Adam, ?eslovas Venclovas, Krzysztof Fidelis, and Burkhard Rost. 1999. "A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment." *Proteins: Structure, Function, and Genetics* 34 (2): 220–23. https://doi.org/10.1002/(sici)1097-0134(19990201)34:2%3C220::aid-prot7%3E3.0.co;2-k.

Zeng, Hongkui. 2022. "What Is a Cell Type and How to Define It?" *Cell* 185 (15): 2739–55. https://doi.org/10.1016/j.cell.2022.06.031.

Zeng, Wen-Feng, Xie-Xuan Zhou, Sander Willems, et al. 2022. "AlphaPeptDeep: A Modular Deep Learning Framework to Predict Peptide Properties for Proteomics." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-34904-3.

Zerbino, Daniel R., and Ewan Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. "ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees." *BMC Bioinformatics* 19 (S6). https://doi.org/10.1186/s12859-018-2129-y.

Zhang, J. 2005. "Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level." *Molecular Biology and Evolution* 22 (12): 2472–79. https://doi.org/10.1093/molbev/msi237.

Zhang, Shuang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. 2023. "Applications of Transformer-Based Language Models in Bioinformatics: A Survey." *Bioinformatics Advances* 3 (1). https://doi.org/10.1093/bioadv/vbad001.

Zhang, Yang, and Jeffrey Skolnick. 2004. "Scoring Function for Automated Assessment of Protein Structure Template Quality." *Proteins: Structure, Function, and Bioinformatics* 57 (4): 702–10. https://doi.org/10.1002/prot.20264.

Zhang, Yang, and Jeffrey Skolnick. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based

on the TM-Score." *Nucleic Acids Research* 33 (7): 2302–9. https://doi.org/10.1093/nar/gki524.

Zhang, Yong, Tao Liu, Clifford A Meyer, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9). https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Zhidian, Hannah K. Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. 2024. "Protein Language Models Learn Evolutionary Statistics of Interacting Sequence Motifs." *Proceedings of the National Academy of Sciences* 121 (45). https://doi.org/10.1073/pnas.2406285121.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2013. "CrossMap: A Versatile Tool for Coordinate Conversion Between Genome Assemblies." *Bioinformatics* 30 (7): 1006–7. https://doi.org/10.1093/bioinformatics/btt730.

Zhao, Qi, Zheng Zhao, Xiaoya Fan, Zhengwei Yuan, Qian Mao, and Yudong Yao. 2021. "Review of Machine Learning Methods for RNA Secondary Structure Prediction." *PLOS Computational Biology* 17 (8): e1009291. https://doi.org/10.1371/journal.pcbi.1009291.

Zhen, Ying, Matthew L. Aardema, Edgar M. Medina, Molly Schumer, and Peter Andolfatto. 2012. "Parallel Molecular Evolution in an Herbivore Community." *Science* 337 (6102): 1634–37. https://doi.org/10.1126/science.1226630.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (1). https://doi.org/10.1038/ncomms14049.

Zhou, Jian, and Olga G Troyanskaya. 2015. "Predicting Effects of Noncoding Variants with Deep Learning–Based Sequence Model." *Nature Methods* 12 (10): 931–34. https://doi.org/10.1038/nmeth.3547.

Zhou, Naihui, Yuxiang Jiang, Timothy R. Bergquist, et al. 2019. "The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes Through Experimental Screens." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1835-8.

Zhu, Wensi, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. 2023. "Evaluation of AlphaFold-Multimer Prediction on Multi-Chain Protein Complexes." *Bioinformatics* 39 (7). https://doi.org/10.1093/bioinformatics/btad424.

Zielezinski, Andrzej, Hani Z. Girgis, Guillaume Bernard, et al. 2019. "Benchmarking of Alignment-Free Sequence Comparison Methods." *Genome Biology* 20 (1). https://doi.org/10.1186/s13059-019-1755-7.

Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. 2017. "Alignment-Free Sequence Comparison: Benefits, Applications, and Tools." *Genome Biology* 18 (1). https://doi.org/10.1186/s13059-017-1319-7.

Zielinska, Dorota F., Florian Gnad, Jacek R. Wiśniewski, and Matthias Mann. 2010. "Precision Mapping of an in Vivo n-Glycoproteome Reveals Rigid Topological and Sequence Constraints." *Cell* 141 (5): 897–907. https://doi.org/10.1016/j.cell.2010.04.012.

Ziemann, Mark, Yotam Eren, and Assam El-Osta. 2016. "Gene Name Errors Are Widespread in the Scientific Literature." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-1044-7.

Zimmerman, Kip D., Mark A. Espeland, and Carl D. Langefeld. 2021. "A Practical Solution to Pseudoreplication Bias in Single-Cell Studies." *Nature Communications* 12 (1). https://doi.org/10.1038/s41467-021-21038-1.

Zook, Justin M., Jennifer McDaniel, Nathan D. Olson, et al. 2019. "An Open Resource for Accurately Benchmarking Small Variant and Reference Calls." *Nature Biotechnology* 37 (5): 561–66. https://doi.org/10.1038/s41587-019-0074-6.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, Zhengting, and Jianzhi Zhang. 2015. "Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations?" *Molecular Biology and Evolution* 32 (8): 2085–96. https://doi.org/10.1093/molbev/msv091.

Zuker, Michael, and Patrick Stiegler. 1981. "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information." *Nucleic Acids Research* 9 (1): 133–48. https://doi.org/10.1093/nar/9.1.133.

Zvelebil, Marketa, and Jeremy O. Baum. 2008. *Understanding Bioinformatics*. Garland Science.

Zverinova, Stepanka, and Victor Guryev. 2021. "Variant Calling: Considerations, Practices, and Developments." *Human Mutation* 43 (8): 976–85. https://doi.org/10.1002/humu.24311.

A Recommended reading

For each chapter, about six papers are recommended for reading in full. Most of them come from the reading lists used for the exam in earlier years. They are listed here with their titles, abstracts and links. In the PDF version the open-access papers follow each list in full; the others are available with your SU login through the library link.

A.1 Sequence searching

Recommended reading for Chapter 1 (listed in Section 1.13).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Smith and Waterman (1981) — The local alignment algorithm, in three pages.
  DOI · via SU library (SU login)

  Identification of common molecular subsequences. Smith & Waterman (1981). *Journal of Molecular Biology*. doi:10.1016/0022-2836(81)90087-5

- Altschul et al. (1997) — Gapped BLAST and PSI-BLAST; still the standard reference for heuristic search and profiles.
  DOI · via SU library (SU login)

  Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul et al. (1997). *Nucleic Acids Research*. doi:10.1093/nar/25.17.3389

  > The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the BLAST programs to be decreased substantially while enhancing their sensitivity to weak similarities. A new criterion for triggering the extension of word hits, combined with a new heuristic for generating gapped alignments, yields a gapped BLAST program that runs at approximately three times the speed of the original. In addition, a method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as gapped BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. PSI-BLAST is used to uncover several new and interesting members of the BRCT superfamily.

- Eddy (2011) — HMMER3; profile HMM search made as fast as BLAST, with a clear account of the statistics.
  full text below · DOI · open access (CC BY 4.0)

  Accelerated Profile HMM Searches. Eddy (2011). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002195

> Profile hidden Markov models (profile HMMs) and probabilistic inference methods have made important contributions to the theory of sequence database homology search. However, practical use of profile HMM methods has been hindered by the computational expense of existing software implementations. Here I describe an acceleration heuristic for profile HMMs, the "multiple segment Viterbi" (MSV) algorithm. The MSV algorithm computes an optimal sum of multiple ungapped local alignment segments using a striped vector-parallel approach previously described for fast Smith/Waterman alignment. MSV scores follow the same statistical distribution as gapped optimal local alignment scores, allowing rapid evaluation of significance of an MSV score and thus facilitating its use as a heuristic filter. I also describe a 20-fold acceleration of the standard profile HMM Forward/Backward algorithms using a method I call "sparse rescaling". These methods are assembled in a pipeline in which high-scoring MSV hits are passed on for reanalysis with the full HMM Forward/Backward algorithm. This accelerated pipeline is implemented in the freely available HMMER3 software package. Performance benchmarks show that the use of the heuristic MSV filter sacrifices negligible sensitivity compared to unaccelerated profile HMM searches. HMMER3 is substantially more sensitive and 100- to 1000-fold faster than HMMER2. HMMER3 is now about as fast as BLAST for protein searches.

- Steinegger and Söding (2017) — MMseqs2; how sensitive search scales to billions of sequences. DOI · via SU library (SU login)

  MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Steinegger & Söding (2017). *Nature Biotechnology*. doi:10.1038/nbt.3988

- van Kempen et al. (2023) — Foldseek; searching with structure instead of sequence. full text below · DOI · open access (CC BY 4.0)

  Fast and accurate protein structure search with Foldseek. van Kempen et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-023-01773-0

  > As structure prediction methods are generating millions of publicly available protein structures, searching these databases is becoming a bottleneck. Foldseek aligns the structure of a query protein against a database by describing tertiary amino acid interactions within proteins as sequences over a structural alphabet. Foldseek decreases computation times by four to five orders of magnitude with 86%, 88% and 133% of the sensitivities of Dali, TM-align and CE, respectively.

The open-access papers are reproduced below under their original licences.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Accelerated Profile HMM Searches

Sean R. Eddy *

HHMI Janelia Farm Research Campus, Ashburn, Virginia, United States of America

Abstract

Profile hidden Markov models (profile HMMs) and probabilistic inference methods have made important contributions to the theory of sequence database homology search. However, practical use of profile HMM methods has been hindered by the computational expense of existing software implementations. Here I describe an acceleration heuristic for profile HMMs, the "multiple segment Viterbi" (MSV) algorithm. The MSV algorithm computes an optimal sum of multiple ungapped local alignment segments using a striped vector-parallel approach previously described for fast Smith/Waterman alignment. MSV scores follow the same statistical distribution as gapped optimal local alignment scores, allowing rapid evaluation of significance of an MSV score and thus facilitating its use as a heuristic filter. I also describe a 20-fold acceleration of the standard profile HMM Forward/Backward algorithms using a method I call "sparse rescaling". These methods are assembled in a pipeline in which high-scoring MSV hits are passed on for reanalysis with the full HMM Forward/Backward algorithm. This accelerated pipeline is implemented in the freely available HMMER3 software package. Performance benchmarks show that the use of the heuristic MSV filter sacrifices negligible sensitivity compared to unaccelerated profile HMM searches. HMMER3 is substantially more sensitive and 100- to 1000-fold faster than HMMER2. HMMER3 is now about as fast as BLAST for protein searches.

Citation: Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7(10): e1002195. doi:10.1371/journal.pcbi.1002195

Editor: William R. Pearson, University of Virginia, United States of America

Received April 27, 2011; **Accepted** July 29, 2011; **Published** October 20, 2011

Copyright: © 2011 Sean R. Eddy. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Howard Hughes Medical Institute (HHMI). HHMI had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: eddys@janelia.hhmi.org

Introduction

Sequence database homology searching is one of the most important applications in computational molecular biology. Genome sequences are being acquired rapidly for an ever-widening array of species. To make maximal use of sequence data, we want to maximize the power of computational sequence comparison tools to detect remote homologies between these sequences, to learn clues to their functions and evolutionary histories. The most widely used tool for sequence comparison and database search is BLAST [1–3].

Since BLAST's introduction, some important advances have been made in the theory of sequence comparison, particularly by using probabilistic inference methods based on profile hidden Markov models (profile HMMs) [4]. Probabilistic modeling approaches provide a consistent framework for parameterizing complex position-specific models of sequence conservation and evolution [5]. Numerous improvements have been made in BLAST in light of these advances [6–9]. Fundamentally, though, the BLAST implementation computes optimal local alignment scores using *ad hoc* gap penalties. This implementation core may not be readily adaptable to a probabilistic insertion/deletion model and the more powerful "Forward/Backward" HMM algorithm that computes not just one best-scoring alignment, but a sum of probabilities over the entire local alignment ensemble. The Forward algorithm allows a more powerful and formal log-likelihood score statistic to be assigned to each target sequence, and Forward/Backward allows confidence values to be assigned to each aligned residue.

Nonetheless, regardless of any of the attractive advantages of HMMs, no implementation of fully probabilistic sequence comparison methods has yet approached the utility of BLAST. The most widely used implementations of profile HMM technology, including HMMER from my laboratory, have been slow and computationally expensive, on the order of 100- to 1000-fold slower than BLAST for a comparably sized search. In an era of enormous sequence databases, this speed disadvantage outweighs any advantage of HMM methods. Profile HMM methods have become important only in the niche of protein domain family analysis, where the speed differential is compensated by being able to use a single profile HMM to represent a family of hundreds of homologous individual sequences [10,11].

HMMER has been a target of many acceleration and optimization efforts [12–15] but these efforts have had limited impact. The only accelerations that have reported large gains have implemented HMMER's native dynamic programming algorithms on specialized hardware, including FPGAs (field-programmable gate arrays) [16–19], VLSI ASICs (special-purpose chips), GP-GPUs (general purpose graphics processor units) [20,21], and large multiprocessor clusters [22,23]. Fewer efforts have been made to develop fast heuristic profile HMM algorithms for standard commodity processors [24–26] in ways comparable to how BLAST heuristically approximates and accelerates Smith/Waterman optimal dynamic programming alignment [27]. The challenge is that to preserve the significant yet narrow gain in sensitivity that profile HMM methods show over BLAST [28–30], any useful profile HMM acceleration heuristic must be more sensitive than BLAST's already excellent heuristics.

Another reason for the limited impact of previous acceleration efforts is that they have almost exclusively focused on accelerating the optimal local alignment scoring algorithm (known as the

Author Summary

Searching sequence databases is one of the most important applications in computational molecular biology. The main workhorse in the field is the BLAST suite of programs. Since the introduction of BLAST in the 1990's, important theoretical advances in homology search methodology have been made using probabilistic inference methods and hidden Markov models (HMMs). However, previous software implementations of these newer probabilistic methods were slower than BLAST by about 100-fold. This hindered their utility, because computation speed is so critical with the rapidly increasing size of modern sequence databases. Here I describe the acceleration methods I implemented in a new, freely available profile HMM software package, HMMER3. HMMER3 makes profile HMM searches about as fast as BLAST, while retaining the power of using probabilistic inference technology.

Viterbi algorithm in the HMM literature) as opposed to the more desirable Forward algorithm. In part, this is because optimal local alignment algorithms are more well known, and in part it is because previous versions of HMMER itself implemented Viterbi rather than Forward scoring. Forward implementations are about 3- to 9-fold slower than Viterbi implementations, and the expected statistical distribution of Forward scores for profile HMMs was not understood well enough to assign accurate E-values (expectation values). I recently described a satisfactory solution to the latter problem [31], which leaves the problem of acceleration.

Here I describe the heuristic acceleration pipeline implemented by HMMER3, a reimplemented version of the HMMER software. In comparison to the previous version of HMMER, HMMER3 is about 100-fold faster because of the use of a new heuristic algorithm called the MSV filter, while also being significantly more powerful because it moves from optimal local Viterbi alignment to full Forward/Backward evaluation of alignment ensembles, exploiting more of the mathematical advantages of probabilistic modeling. Thus HMMER3 is now about as fast as BLAST, while extending the performance advantages of profile HMM methods.

Results

Overview

The main algorithm that accelerates HMMER3 is called MSV, for Multiple (local, ungapped) Segment Viterbi. It was inspired by a technique used in ParAlign [32]. As shown in Figure 1, the MSV model is an ungapped version of HMMER3's multihit local alignment model. MSV's probabilistic model of multihit ungapped local alignment is achieved simply by ignoring the match, delete, and insert state transitions of the original profile and implicitly treating match-match transitions as 1.0.

An MSV score is essentially analogous to BLAST's "sum score" of one or more ungapped HSPs (high scoring pairs). A difference is that MSV does not impose alignment consistency (two ungapped alignments are not required to be consistent with a single gapped alignment). In a filtering heuristic, this difference is not important. HMMER3 calculates the MSV score directly by dynamic programming, bypassing the word hit and hit extension heuristics of BLAST.

The fact that MSV essentially bypasses two of BLAST's main heuristics provides an intuitive argument why MSV scores are expected to be a more sensitive overall heuristic than BLAST's approach. However, I have not attempted to rigorously compare the performance of HMMER's MSV heuristic to other acceleration heuristics such as those in BLAST or FASTA.

The HMMER3 implementation takes advantage of several synergistic statistical and computational features of the MSV model. I summarize these features here before describing them in detail:

- MSV alignment scores can be calculated efficiently using so-called "striped" vector-parallel techniques originally developed for Smith/Waterman local sequence alignment [33], because the MSV model removes deletion and insertion states that interfere with vector parallelism.
- Because the MSV model gives predictable score distributions for nonhomologous sequences, with scores confined to a narrow range that is largely independent of query and target sequence characteristics, MSV values can be approximated with reduced precision (8 bits, in a score range of 0–255). This allows a 16-fold vector parallelism in current commodity processors with 128-bit vector registers.
- The MSV model remains a full probabilistic local alignment model, so MSV scores obey conjectures about the expected Gumbel distribution of probabilistic local alignment scores [31]. This allows the rapid calculation of P-values.
- Because we can calculate MSV P-values, we can use MSV scores as a tunable and selective sequence filter. If a target sequence has an MSV score with a P-value less than a chosen threshold, we pass the entire sequence to more accurate and computationally intensive scoring algorithms. By definition, the P-value threshold is the fraction of nonhomologous sequences expected to pass the filter.

The MSV filter is a heuristic acceleration, not guaranteed to find all high-scoring targets. Overall performance of the HMMER3 acceleration pipeline in terms of speed, specificity, and sensitivity depends on several issues and tradeoffs, including how fast the filters are, how accurately and quickly P-values can be estimated for filter scores, and whether a threshold on MSV P-values can be set to remove most nonhomologs while removing few if any true homologs that an unfiltered search would have detected. These are empirical questions, which I have addressed by benchmarking experiments.

The following sections, especially on vector parallelization and on assuring that scores can be kept in limited numeric ranges, are necessarily technical and terse. On a first reading, the reader may want to skip or skim ahead to the "HMMER3 acceleration pipeline" section to see how these technical aspects fit together into an overall scheme, and how that acceleration scheme performs.

MSV model: notation and parameterization

The MSV score for target sequence $\mathbf{x}$ is a standard HMM Viterbi score, a log likelihood ratio score of a single optimal (maximally likely) alignment: the ratio of the probability of the optimal alignment $\mathring{\pi}$ for $\mathbf{x}$ given the MSV model M^{MSV} and the probability of the sequence given a null hypothesis model R:

$$\mathrm{S}^{\mathrm{MSV}}(\mathbf{x}) = \log_2 \frac{\mathrm{Prob}(\mathbf{x},\mathring{\pi}|M^{\mathrm{MSV}})}{\mathrm{Prob}(\mathbf{x}|R)}$$

For a query of length M positions, the MSV profile has KM match emission parameters (where K is the alphabet size, 4 nucleotides or 20 amino acids), plus $M+8$ additional state transition parameters involving the flanking N, B, E, C, and J states that account for nonhomologous residues. Other state

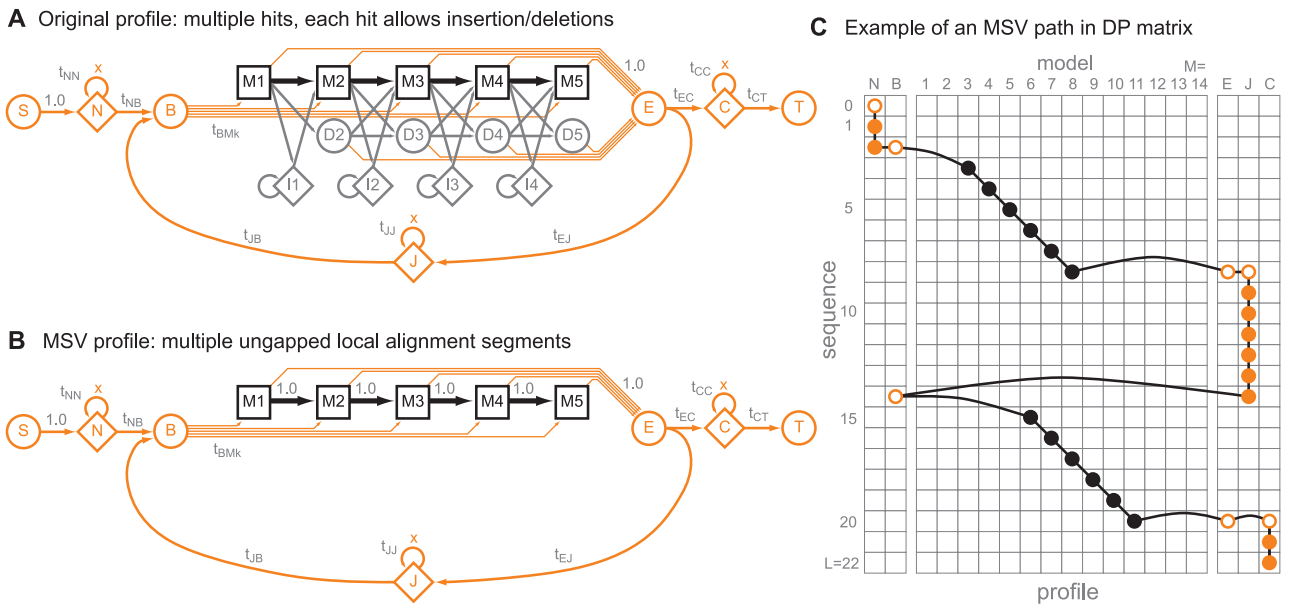


Figure 1. The MSV profile. A: Profile HMM architecture used by HMMER3 [4,5,31]. Regions homologously aligned to the query are represented by a linear core model consisting of M consensus positions (in this example, $M=5$), each consisting of a match, a delete, and an insert state (shown as boxes marked M, circles marked D, and diamonds marked I), connected by state transition probabilities (arrows). Match states carry position-specific emission probabilities for scoring residues at each consensus position. Insert states emit residues with emission probabilities identical to a background distribution. Additional flanking states (marked N, C, and J) emit zero or more residues from the background distribution, modeling nonhomologous regions preceding, following, or joining homologous regions aligned to the core model. Start (S), begin (B), end (E) and termination (T) states do not emit. B: The MSV profile is formed by implicitly treating all match-match transition probabilities as 1.0. This corresponds to the virtual removal of the delete and insert states. The rest of the profile parameterization stays the same. This model generates sequences containing one or more ungapped local alignment segments. Note that both models appear to be improperly normalized; for example, each match state in the MSV model has probability 1.0 local exit transition (orange arrows) in addition to the probability 1.0 match-match transition. This is because of a trick used to establish a uniform local fragment length distribution, in which these profiles are collapsed representations of a much larger (and properly normalized) "implicit probability model", as explained in [31]. C: An example of what an alignment of a larger MSV profile (of length $M=14$) to a target sequence (of length $L=22$) might look like, as a path through a dynamic programming (DP) matrix. Here, the model identifies two high-scoring ungapped alignment segments (black dots, indicating residues aligned to profile match states), and assigns all other residues to N, J, and C states in the model (orange dots; unfilled indicates a "mute" nonemitting state or state transition). Note that the ungapped diagonals are not enforced to be consistent with a single gapped alignment.
doi:10.1371/journal.pcbi.1002195.g001

transitions in the original profile are ignored, which means implicitly treating match-match transitions as 1.0.

The null model R is assumed to be an HMM with a single state R emitting residues a with background frequencies $f(a)$ (i.e. a standard i.i.d. null model: independent, identically distributed residues), with a geometric length distribution specified by a transition parameter t_{RR}.

The KM position-specific match scores $\sigma_k(a)$ are precomputed as log-odds ratios for a residue a emitted from match state M_k with emission probability $e_k(a)$, compared to the null model background frequencies f_a:

$$\sigma_k(a) = \log_2 \frac{e_k(a)}{f_a}.$$

These match scores (as well as the emission probabilities and background frequencies) are the same as in the original profile.

The only state transition parameters in the MSV model are those that control target sequence length modeling, the uniform local alignment fragment length distribution, and the number of hits to the core homology model per target sequence [31]. These too are identical to the parameterization of the original profile [31]. Specifically, they are set as follows for a target sequence of length L residues and a model of length M consensus positions:

Target sequence length modeling:

$$t_{NN} = t_{CC} = t_{JJ} = \frac{L}{L+3},$$
$$t_{NB} = t_{CT} = t_{JB} = \frac{3}{L+3},$$
$$t_{RR} = \frac{L}{L+1}.$$

Uniform local alignment fragment length distribution:

$$t_{BM_k} = \frac{2}{M(M+1)},$$
$$t_{M_kE} = 1.0.$$

Multiple hits per target:

$$t_{EC} = 0.5,$$
$$t_{EJ} = 0.5.$$

MSV score algorithm (serial version)

The MSV alignment score can be calculated by a dynamic programming recursion in a two-dimensional matrix $W(i,s)$ indexed by HMM state (M_k,*NBECJ*) and target sequence residue i:

Initialization:

$$W(0,\mathrm{E}) = W(0,\mathrm{C}) = W(0,\mathrm{J}) = -\infty;$$
$$W(0,\mathrm{M}_k) = -\infty, \forall k;$$
$$W(i,\mathrm{M}_0) = -\infty, \forall i;$$
$$W(0,\mathrm{N}) = 0;$$
$$W(0,\mathrm{B}) = t_{\mathrm{NB}}.$$

Recursion:

for $i = 1$ to L:

for $k = 1$ to M:

$$W(i,\mathrm{M}_k) = \sigma_k(x_i) + \max\left\{ W(i-1,\mathrm{M}_{k-1}),\ W(i-1,\mathrm{B}) + \log_2 t_{\mathrm{BM}_k} \right\};$$

$$W(i,\mathrm{E}) = \max_k W(i\mathrm{M}_k);$$

$$W(i,\mathrm{N}) = W(i-1,\mathrm{N}) + \log_2\ t_{\mathrm{NN}};$$

$$W(i,\mathrm{J}) = \max\{ W(i-1,\mathrm{J}) + \log_2\ t_{\mathrm{JJ}},\ W(i-1,\mathrm{E}) + \log_2\ t_{\mathrm{EJ}} \};$$

$$W(i,\mathrm{C}) = \max\{ W(i-1,\mathrm{C}) + \log_2\ t_{\mathrm{CC}},\ W(i,\mathrm{E}) + \log_2\ t_{\mathrm{EC}} \};$$

$$W(i,B) = \max\{ W(i-1,\mathrm{N}) + \log_2\ t_{\mathrm{NB}},\ W(i,J) + \log_2\ t_{\mathrm{JB}} \}.$$

Termination: $S^{\mathrm{MSV}}(\mathbf{x}) = W(L,\mathrm{C}) + \log_2 t_{\mathrm{CT}} - (L \log_2 t_{\mathrm{RR}} + \log_2(1 - t_{\mathrm{RR}}))$

The M_0 "state" in the initialization is solely needed for a boundary condition; there is no such state in the model.

Log-odds ratio scoring relative to the null model is built into the calculation, in the match scores $\sigma_k(a)$ and in counting the total (constant) null model state transition contribution of $L \log_2\ t_{\mathrm{RR}} + \log_2(1 - t_{\mathrm{RR}})$ as terms in the DP termination step.

Like other linear sequence alignment recursions, the algorithm requires $O(ML)$ time. It is implemented in a single lattice row of $O(M)$ space for purposes of obtaining just the optimal score. In the HMMER3 source code, this algorithm is implemented in generic_msv.c::p7_GMSV().

MSV score algorithm: SIMD vector parallelization

The MSV algorithm is highly amenable to vector parallelization using commodity SIMD (single instruction, multiple data) instructions, such as the Streaming SIMD Extensions (SSE) instructions on Intel-compatible systems and Altivec/VMX instructions on PowerPC systems. These vector instruction sets use 128-bit vectors to compute up to 16 simultaneous operations.

Several vector methods have been described for accelerating classical Smith/Waterman local sequence alignment [34,35], and methods for accelerating Smith/Waterman dynamic programming (DP) recursions are readily adapted to profile HMMs. A remarkably efficient vector-parallel approach called *striped* Smith/Waterman was described by Farrar [33].

Striping addresses a challenge in the data dependency pattern in Smith/Waterman-style dynamic programming recursions. The calculation of each cell (i,k) in the dynamic programming lattice requires having previously calculated cells $(i-1,k)$, $(i,k-1)$, and $(i-1,k-1)$. In a row-vectorized implementation, V individual cells (typically 4, 8, or 16) are stored in each individual vector, such that each row i of the vectorized DP matrix stores cells $k = 1..M$ in Q vectors numbered $q = 1..Q$, where $Q = (M + V - 1)/V$. In Farrar's approach, cells $k = 1..M$ are assigned nonconsecutively to vectors $q = 1..Q$ in a striped pattern (Figure 2). In striped vectors, when we calculate the set of several cells (i,k) contained in one vector (i,q) on a current row, all the previous diagonal cells $(i-1,k-1)$ that we need are neatly available in the correct order in a vector $(i-1,q-1)$ on the previous row, and the cells above are in vector $(i-1,q)$. Striping minimizes expensive operations such as shifting or rearranging cell values inside vectors. The disadvantage is that calculations on delete paths (dependent on cells $i,k-1$ to the left) may need to be fully serialized. Farrar described effective techniques for minimizing this problem. In the MSV algorithm, because only ungapped diagonals are calculated, this drawback is avoided altogether. The essential idea of how striped indexing works is schematized in Figure 2.

To maximize parallelism, I implemented MSV as a 16-fold parallel calculation with score values stored as 8-bit unsigned integers restricted to range 0..255. This takes advantage of the fact that local alignment scores under HMMER3's probabilistic model have a narrow and predictable dynamic range, enabling a numerical stability analysis that justifies using reduced precision. (The details of this analysis are given in the next section.) This rescaling is specified by three values (base, bias, and scale), where "base" is an initial offset from zero to make MSV scores nonnegative (default: 190), "scale" is the scaling factor (default 3, so MSV scores are in units of one-third bits), and "bias" is an offset on individual residue scores, used to make all individual residue scores unsigned byte costs relative to the maximum residue score. Using the scale and bias terms, position-specific residue scores $\sigma_k(x_i)$ are converted to precomputed scaled costs by $-1 * \mathrm{roundf}(\mathrm{scale} * \sigma_k(x_i)) + \mathrm{bias}$ (saturated at a maximum cost of 255) and stored in striped order in vectors $\hat{\sigma}_{x_i}(q)$ (Figure 2). Transition scores $\log_2\ t_{xx}$ are converted to precomputed scaled costs $\hat{t}_{xx}$ by $-1 * \mathrm{roundf}(\mathrm{scale} * \log_2\ t_{xx})$ (saturated at a maximum cost of 255).

To define MSV's SIMD recursion, I will use five pseudocode vector instructions for operations on $b-$bit integers ($b = 8$ in our implementation), either scalars x or vectors v containing V $b-$bit integer elements numbered $v[1]..v[z]..v[V]$. Each of these operations are either available or easily constructed in both SSE and Altivec/VMX:

Operation	Pseudocode	Definition
saturated addition	v = vec_adds(v1,v2)	$v[z] = \mathrm{MIN}(2^b - 1, v1[z] + v2[z])\ \forall z,$
saturated subtraction	v = vec_subs(v1,v2)	$v[z] = \mathrm{MAX}(0, v1[z] - v2[z])\ \forall\ z,$
max	v = vec_max(v1,v2)	$v[z] = \mathrm{MAX}(v1[z], v2[z])\ \forall\ z,$
assignment	v = vec_splat(x)	$v[z] = x\ \forall\ z,$
right shift	v = vec_rightshift(v)	$v[1] = 0;\ v[z] = v[z\text{-}1]\ \forall\ z > 1.$
horizontal max	x = vec_hmax(v)	$x = \max_z v[z]$

In this pseudocode, the vectorized MSV algorithm is the following:

Initialization:

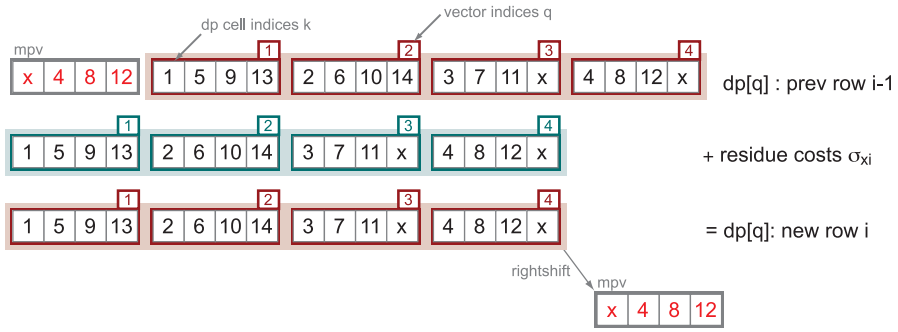


Figure 2. Illustration of striped indexing for SIMD vector calculations. The top row (magenta outline) shows one row of the dynamic programming lattice for a model of length $M=14$. Assuming an example of vectors containing $V=4$ cells each, the 14 cells $k=1..14$ are contained in $Q=4$ vectors numbered $q=1..4$. (Two unused cells, marked x, are set to a sentinel value.) In the dynamic programming recursion, when we calculate each new cell k in a new row i, we access the value in cell $k-1$ in the previous row $i-1$. With striped indexing, vector $q-1$ contains exactly the four $k-1$ cells needed to calculate the four cells k in a new vector q on a new row of the dynamic programming matrix (turquoise outline). For example, when we calculate cells $k=(2,6,10,14)$ in vector $q=2$, we access the previous row's vector $q-1=1$ which contains the cells we need in the order we need them, $k-1=(1,5,9,13)$ (dashed lines and box). If instead we indexed cells into vectors in the obvious way, in linear order ($k=1..4$ in vector $q=1$ and so on), there is no such correspondence of $q,q-1$ with four $k-1,k$'s, and each calculation of a new vector q would require expensive meddling with the order of cells in the previous row's vectors. With striped indexing, only one shift operation is needed per row, outside the innermost loop: the last vector on each finished row is rightshifted (mpv, in grey with red cell k indices) and used to initialize the next row calculation.
doi:10.1371/journal.pcbi.1002195.g002

$$\text{biasv} = \text{vec\_splat}(\text{bias})$$
$$\text{dp[q]} = \text{vec\_splat}(0)\ \forall q$$
$$\text{xB} = \text{base} - \hat{t}_{NB}$$
$$\text{xJ} = 0$$

Recursion:

for $i=1$ to L:

$$\text{xEv} = \text{vec\_splat}(0)$$
$$\text{xBv} = \text{vec\_splat}(\text{xB} - \hat{t}_{BM})$$
$$\text{mpv} = \text{vec\_rightshift}(\text{dp}[\text{Q}-1])$$

for $q=1$ to Q:

$$\text{tmpv} = \text{vec\_max}(\text{mpv}, \text{xBv})$$
$$\text{tmpv} = \text{vec\_adds}(\text{tmpv}, \text{biasv})$$
$$\text{tmpv} = \text{vec\_subs}(\text{tmpv}, \hat{\sigma}_{x_i}(q))$$
$$\text{xEv} = \text{vec\_max}(\text{xEv}, \text{tmpv})$$
$$\text{mpv} = \text{dp[q]}$$
$$\text{dp[q]} = \text{tmpv}$$

$$\text{xE} = \text{vec\_hmax}(\text{xEv})$$
$$\text{xJ} = \max\{\text{xJ}, \text{xE} - \hat{t}_{EJ}\}$$
$$\text{xB} = \max\{\text{base}, \text{xJ}\} - \hat{t}_{JB}$$

Termination: $S^{\text{MSV}}(\mathbf{x}) \simeq \dfrac{\text{xJ} - \hat{t}_{CT} - \text{base}}{\text{scale}} - 4.3 - (L \log_2 t_{\text{RR}} + \log_2 t_{\text{RT}})$

The constant term of -4.3 bits in the termination step arises from an approximation that deals with roundoff error in counting t_{NN} and t_{CC} transition costs. This is explained in the following section. The termination condition is assuming that $\hat{t}_{EC} = \hat{t}_{EJ}$, so that values for the C state are the same as for the J state (thus saving having to calculate C state values in the recursion).

This algorithm is implemented both for SSE and Altivec/VMX instructions in the HMMER3 source code in impl_{sse,vmx}/msvfilter.c::p7_MSVFilter().

Analysis of consequences of reduced numerical precision

This section is particularly technical, and may be skipped in a first reading. In reducing the dynamic range of score calculations to small unsigned integers, we must make sure that underflow or overflow either do not occur, or have no erroneous consequences. We must also be sure that the magnitude of any accumulated roundoff error is tolerable. Because the HMMER3 acceleration pipeline (described below) uses vector-parallel, striped, reduced precision implementations of both the MSV algorithm (described above) and the standard Viterbi (optimal alignment) algorithm for the original profile model with insertions/deletions, the following analysis considers both MSV and Viterbi scores.

For underflow, we use the fact that there is a lower bound on optimal local alignment scores as a function of model length M and target sequence length L. In the worst possible positive-scoring optimal local alignment, the core profile matches only one match state M_k against one residue x_i with a score $\sigma_k(x_i) \geq 0$, and the remaining $L-1$ residues of the target sequence are accounted for by flanking N and C states. The worst case therefore has a

log-odds likelihood score S of

$$S \geq (i-1)\log_2 t_{NN} + \log_2 t_{NB} + \log_2\ t_{BM_k} + \log_2\ t_{M_kE} + \log_2\ t_{EC} + (L-i)\log_2\ t_{CC} + \log_2\ t_{CT} - L\log_2\ t_{RR} - \log_2\ t_{RT}$$

which is:

$$S \geq 2\log_2 \frac{3}{L+3} + (L-1)\log_2 \frac{L}{L+3} + \log_2 \frac{2}{M(M+1)} + \log_2 \frac{1}{2} - L\log_2 \frac{L}{L+1} - \log_2 \frac{1}{L+1} \text{ bits.}$$

Known protein sequences can be over 30,000 residues long (human titin, for example, is 34,350aa). If we specify $L,M \leq 10^6$ as design limits, we can assume a lower score bound of $S \geq -60$ bits for optimal local alignments.

For managing overflow, we use the fact that we will only use a reduced-precision implementation as a filter on target sequences. Any sequence with a P-value $\leq$ a chosen threshold will be passed on to a slower routine for recomputation at full precision. Using *saturated* arithmetic instructions, any target sequences that overflow will be scored as the highest possible score. Now we only need to be able to guarantee that the upper score bound has a P-value $\leq$ the lowest P-value threshold we ever plan to use. From Milosajevic [31,36], we know a conservative bound $P(S > t) \leq 2^{-t}$ for a bit score threshold t. For a design limit allowing filter thresholds $\geq 10^{-5}$, an upper score bound of 17 bits suffices. (0.02 is the default P-value threshold for MSV, and 0.001 is the default for Viterbi scores, as discussed below).

This range of -60 to 17 bits applies to *complete* optimal local alignments; in individual cells of the dynamic programming calculation, we need a little more dynamic range. A high scoring alignment of 17 bits, for example, will have a score of more than 17 bits in the last cell that aligns a match state to a homologous residue, because this state is always followed by negative scores from EC, CC, and CT transitions in the optimal alignment. Taking this into account (including some order of evaluation issues - the fact that the contributions of some transitions, including the null model's contributions, are included in a termination step after the dynamic programming recursion is complete) it can be shown that a range of $-61 \ldots 21$ bits suffices to guarantee that no DP cell involved in an optimal local alignment of range $-60 \ldots 17$ bits will underflow or overflow.

These same bounds apply to both the original local alignment model (Viterbi alignments with insertions and deletions) and the MSV model, because no step in ascertaining these bounds required any consideration of the transition probabilities in the core model (match, insertion, and delete states).

Thus we need a dynamic range of 82 bits (-61 … 21 bits), and the maximum range of an 8-bit integer is 256 values, so scaling log-odds scores to units of 1/3 bits suffices. (Coincidentally, this is comparable to the scaling and roundoff of standard scoring matrices used by BLAST or FASTA; BLOSUM45, for example, is in units of 1/3 bits.) A "base" offset term is then used to adjust the represented value range to the range of bit scores. For unsigned 8-bit integers, a base of +190 means that values 0..255 represent the range of $\frac{-190}{3} \ldots \frac{65}{3}$ bits.

Rounding scores to the nearest 1/3 bit introduces a roundoff error of $-\frac{1}{6} \ldots \frac{1}{6}$ bit per scoring term. A sum of N independent, identically distributed random deviates uniformly distributed on an interval (a,b) has mean zero and variance $N\frac{(b-a)^2}{12}$. Because a local alignment score for a target sequence of length L is modeled as a sum of $\geq 2L$ emission and transition scoring terms, even if each term's roundoff error were independent and uniformly distributed, accumulated roundoff error would be large (normally distributed with mean zero and variance $\geq 2L\frac{(1/3)^2}{12}$; so for $L = 400$ the accumulated error would have a standard deviation of ± 2.7 bits). Worse, roundoff errors are neither independent nor uniformly distributed. A particularly bad case is contributed by transition probabilities t close to 1.0, such as most match-match transitions in the original gapped profile model, where $3\log_2\ t$ for all sufficiently large t rounds to a zero cost. Another bad case is contributed by HMM states that have self-loops, such as insert-insert transitions, where a roundoff error is multiplied by the number of times the state is visited. These two bad cases make the self-loops producing chains of N, C, J, or insert states particularly problematic, because these self-transition probabilities are often close to 1.0; an entire chain of them often gets scored as zero, accumulating a large roundoff error.

The MSV model and its implementation use several features to reduce roundoff error to tolerable limits. First, by eliminating match, delete, and insert transitions and setting all match-match transition probabilities to 1.0 (thus zero cost), the MSV model itself has already eliminated many of the transitions that accumulate non-independent roundoff error, leaving in the core model only the M match state emission probabilities (which are all approximately independent and uncorrelated as far as roundoff error analysis is concerned). Second, the emission probabilities in N, C, and J states are assumed to be equal to the background (null model) frequencies, so the emission scores in N, C, and J are treated as zero by construction, thus they contribute no roundoff error terms. Third, we can take advantage of the fact that the total contribution of the NN, CC, and JJ transitions approximates a constant for sufficiently large L, because a local alignment typically assigns nearly all residues of the target sequence to N, C, and J states and few to match states. Thus we expect a typical local alignment to involve on the order of L NN, CC, and JJ transitions, each scoring $\log_2 \frac{L}{L+3}$, and $L\log_2 \frac{L}{L+3} \simeq -4.3$ bits for large L. Therefore we can score NN, CC, and JJ transitions as zero cost during the recursion, then later add a constant -4.3 bits back onto the score to approximate their missing contribution. This approach may alter the optimal local alignment (during the recursion, paths using NN, CC, and JJ transitions look more favorable than they actually are) but in a score filter, we are not interested in the optimal alignment, only its score.

Thus the roundoff error in a reduced-precision MSV algorithm implementation consists of a bias arising from treating the NN, CC, JJ contributions totalling $(L-\ell)\log_2 \frac{L}{L+3} \geq L\log_2 \frac{L}{L+3} \geq -4.3$ bits as a constant -4.3 bits, and a sum of more or less independent and uniformly distributed error terms including five or more log transition probabilities (t_{NB},t_{BM_k},t_{EC},t_{JB},t_{CT}) and ℓ emission scores for match states involved in ungapped alignment diagonals of total length ℓ residues. For large L, and assuming $\ell \simeq 40$ or so for typical MSV alignments at the edge of statistical significance, a back of the envelope calculation suggests an expected error of mean zero and standard deviation of about $\sqrt{\ell\frac{(1/3)^2}{12}} \simeq 0.6$ bits, and a worst-case maximum error of about $\ell/6 \simeq \pm 6$ bits. Because higher-scoring alignments involve more match emission terms than low-scoring alignments, ℓ is correlated

to the optimal score, so the roughly Gaussian error distribution for a given ℓ will be convolved with a Gumbel score distribution, resulting in a slightly non-Gaussian error distribution with some skew towards the higher error side.

To confirm that this expected roundoff error agrees with empirical observation, I performed simulations in which I examined the differences in MSV scores of a reduced precision implementation (unsigned bytes in $\frac{1}{3}$ bit units) compared to a full-precision floating point implementation. I did this for many different profiles (9,318 models from Pfam release 22) aligned to 1,000 random sequences of varying lengths $L=25,100,400,1600,6400,25600$, and to 207,132 real sequences in SwissProt 49.0 (in UniProt 7.0). This experiment showed roundoff errors for each model were distributed with a standard deviation of 0.4–0.6 bits for real sequences in Swissprot and for random sequences of each length. The mean error was approximately zero for random sequences of lengths >100; a 0.4 bit mean underestimate for UniProt sequences; a 0.5 bit mean underestimate for $L=100$ random sequences; and a 2.2 bit mean underestimate for $L=25$ random sequences. On real UniProt sequences, I observed some extreme differences of up to ± 11 bits. These invariably corresponded to long and highly biased composition sequences, where presumably the alignment length ℓ was large, increasing the potential for accumulated roundoff error. Because these were rare, and MSV is to be used only as a filter, these extremes seem safe to ignore. Overall the range of roundoff error appears tolerable, particularly for large L.

MSV scores obey conjectures allowing fast P-value determination

Previously [31] I conjectured that expected scores for certain probabilistic local alignment models, including the HMMER3 local alignment model, follow easily predictable distributions. Specifically, I conjectured that optimal alignment Viterbi bit scores show Gumbel distributions of fixed slope $\lambda \simeq \log 2$, and the high scoring tail of Forward bit scores follow exponentials of the same slope $\lambda \simeq \log 2$. The MSV model is fully probabilistic and thus ought to obey these conjectures. Therefore MSV optimal alignment scores are predicted to follow a Gumbel distribution

$$P(S \geq t) = 1 - \exp\left[-e^{-\lambda(t-\mu)}\right]$$

with slope $\lambda \simeq \log 2$ and a location μ that is estimated by fitting to a small simulation of scores from 200 or so random sequences.

These statistical conjectures are best obeyed by models or scoring systems with high relative entropy per position (i.e. high mean expected score) [31]. Default HMMER3 models have low relative entropy per position (about 0.6 bits/position) because HMMER3 model parameterization uses a technique called entropy-weighting [29,37], an *ad hoc* method to re-weight the effective number of observed sequences relative to the prior (pseudocounts) to achieve a desired relative entropy target. The standard pairwise residue alignment scoring system (BLOSUM62) has a similar relative entropy of about 0.6 bits per aligned position. At lower relative entropy per position, longer alignments are required to achieve high scores and a finite-length "edge effect" becomes considerable. HMMER3 ameliorates edge effect by calculating an *ad hoc* corrected $\hat{\lambda} = \log_2 + \frac{1.44}{MH}$, where M is the length of the profile in match states and H is the relative entropy per match emission state in bits [31]. Although the same *ad hoc* correction suffices for both Viterbi and Forward distributions, it was obtained by empirical fitting with little theoretical guidance, so there is little reason to trust that the same correction would apply to MSV scores. Therefore I empirically tested the ability to estimate accurate P-values for MSV scores for a wide range of HMMER/Pfam models.

Figure 3A shows an example of an MSV score distribution for one typical profile HMM (the CNP1 model from Pfam version 24, representing a lipoprotein family, chosen because it has the median length, median number of representative sequences, and median average pairwise identity over all Pfam 24 seed alignments), for 10^8 scores of random i.i.d. sequences of varying lengths. For all but the shortest sequences (L = 25), the observed score distributions closely match the conjectured distribution including the *ad hoc* edge correction term (orange line).

Figure 3B shows results of simulations in which 11,912 different profile HMMs from Pfam version 24 [11] were scored by the MSV algorithm against 10^5 random sequences of length 400, the resulting distributions were fit to Gumbel distributions to determine maximum likelihood estimates of μ and λ, and a histogram of the λ estimates is plotted. For high relative entropy models (grey line), this distribution is tightly clustered at the expected $\lambda = \log 2$. For default entropy-weighted models (black line), the distribution is broader with a higher mean, in accordance with what is observed for Viterbi scores and attributed to finite-length edge effect.

Figure 3C shows a direct evaluation of the accuracy of MSV P-values across many Pfam models and various random sequence lengths. For each of 11,912 Pfam 24 models, MSV scores are calculated for 500,000 random sequences generated at each of several lengths $L=25,50,100,400,1600,6400$ and 25600, and the number of random sequences that pass the MSV filter at $P \leq 0.02$ is counted. If P-values are accurate, we would expect to see an approximately normal distribution centered at 2% of random sequences passing the filter. Within a tolerance of about 2-fold error, this is true for almost all models and for target sequence lengths ≥ 100 or so. A few models have less well-predicted distributions and produce modest outliers. The largest problems appear with short target sequences ($L=25,50$) where P-values can be up to about five-fold overestimated (i.e., fewer sequences pass than predicted), as seen in the CNP1 example in panel A. Default entropy-weighted models (black) are more affected than models with high relative entropy (gray).

This analysis shows that in general, reasonably accurate P-values for MSV scores can be obtained. It also shows that on short sequences of $L \leq 50$ or so, the MSV filter may be too aggressive (removing more sequences than predicted), and that a few models are outliers with either too few or too many sequences getting through the filter. These are minor issues that would be good to deal with in the future.

Vector parallelization of the Forward and Backward algorithms

The MSV implementation described above is about 500-fold faster than a standard serial implementation of the full Forward algorithm. This means that a search will still be rate-limited by the speed of the computationally intensive Forward/Backward calculations. Suppose we allow the top 2% (1/50) of sequences through the MSV filter to full HMM Forward log-likelihood scoring; then Forward must be no more than 50-fold slower than MSV, or Forward will be rate-limiting. It was therefore necessary to seek significant accelerations of at least an order of magnitude in the implementations of the Forward and Backward algorithms.

Numerical underflow is a problem for implementing the Forward and Backward algorithms. The probability of a partial alignment path generally underflows the smallest representable floating-point value. In a Viterbi implementation, underflow is avoided by working in the log probability domain, replacing

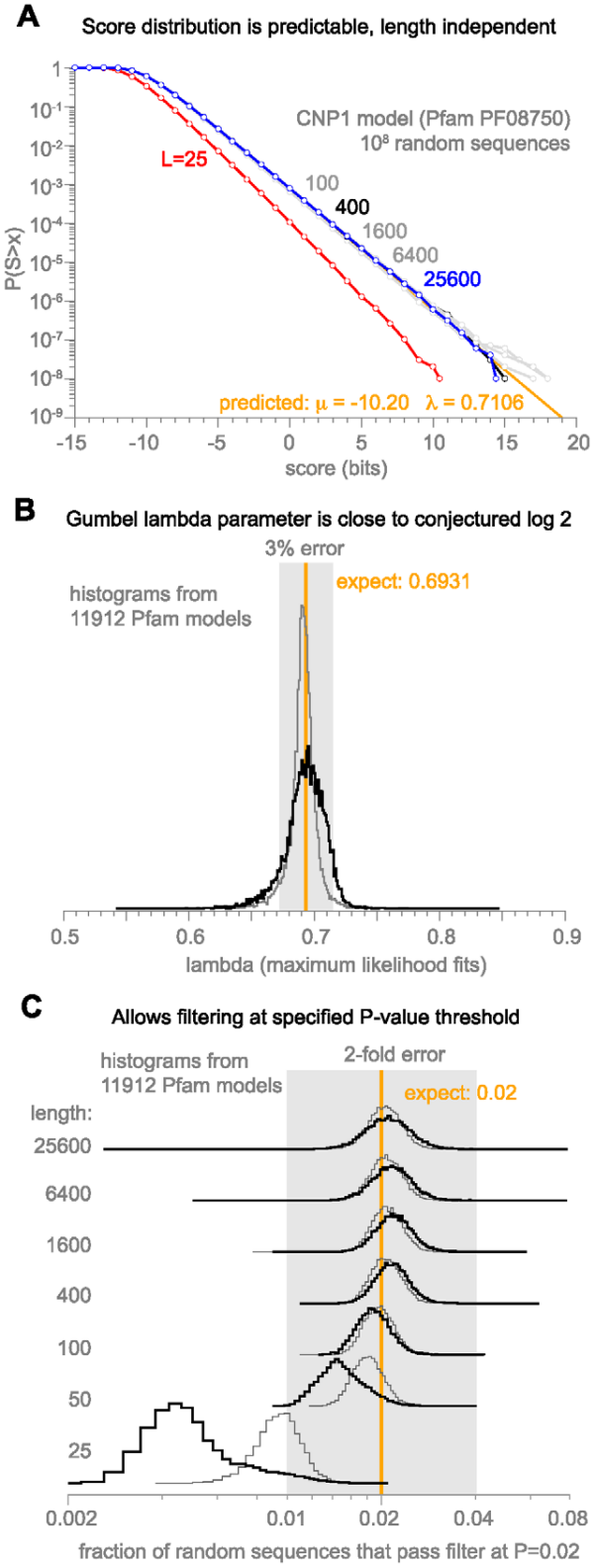


Figure 3. MSV scores follow a predictable distribution. A: example MSV score distributions for a typical Pfam model, CNP1, on 10^8 random i.i.d. sequences of varying lengths from 25 to 25,600, with the shortest, typical, and longest lengths highlighted as red, black, and blue lines, respectively. The predicted distribution, following the procedure of [31] including an edge correction on the slope λ, is shown in orange (though largely obscured by the data lines right on top of it). B: Histogram of maximum likelihood λ values obtained from score distributions of 11,912 Pfam models, showing that most are tolerably close to the conjectured $\lambda = \log_2$, albeit with more dispersion for default entropy-weighted models (black line) than high relative entropy models without entropy-weighting (gray line). C: The observed fraction of nonhomologous sequences that pass the filter at a P-value of 0.02 should be 0.02. Histograms of the actual filter fraction for 11,912 different Pfam 24 models are shown, for a range of random sequence lengths from 25 to 25,600, for both default models (black lines) and high relative entropy models with no entropy weighting (gray lines).
doi:10.1371/journal.pcbi.1002195.g003

multiplication and maximization of probabilities with addition and maximization of their logarithms [5,38]. However, the Forward and Backward dynamic programming recursions require addition of partial paths in the probability domain.

In both sequence analysis and speech recognition HMM applications, this problem is customarily solved by working in the log probability domain and implementing a "log-sum" operation, such that addition $c = a + b$ in the probability domain is replaced by $A = \log a$, $B = \log b$, and $\log c = \log(a+b) = \log(e^A + e^B)$ [5,39]. An efficient log-sum operation rearranges the log-sum to $A + \log(1 + e^{B-A})$ for $A \geq B$, and finds an approximate $\log(1 + e^{B-A})$ term to add to A in a precalculated lookup table indexed by the difference $(A - B)$ scaled and rounded to an appropriate precision. The lookup table has finite size, because the term is negligible for large $(A - B)$, but nonetheless it is large (16,000 entries in the HMMER3 "generic" non-vectorized implementation; see logsum.c). I do not know how to implement a large lookup table efficiently in SIMD vector instructions. Only small lookup tables appear feasible using vector permutation instructions (up to perhaps 256 entries in Altivec/VMX, fewer in SSE).

Another approach is rescaling [5,38]. In rescaling, the entries in each row i (for each target residue x_i) of the dynamic programming matrix are multiplied by some scale value s_i. The scale values s_i are chosen to keep the largest entries in each row within the allowable numeric range. If the smallest values in a row differ from the largest by greater than the numeric range, the smallest values still underflow, but for many HMMs one can show that no partial path prefix that underflows could have ever rebounded to have a non-negligible probability as a complete path. However, in general this is not the case for profile HMMs (nor for other HMMs with paths involving silent states), because of the possibility of long deletions. Even an optimal alignment can contain a long D-D-D path along a single row i. After many DD transition probabilities are multiplied together, the values in the states at the start versus the end of a long deletion path on the same row can differ by more than the allowable range. A standard normalized IEEE754 32-bit float type has a range of about $-2^{128}..2^{128}$, equivalent to 256 bits in the log-odds score domain. Given a typical deletion extension penalty of about -1 to -2 bits, a deletion of about 200 residues or so will typically underflow the rescaled delete states in the correct path. Deletions of this length are rare, but do occur.

I use the following steps to make an approach that I call "sparse rescaling" work for HMMER3's SIMD vector implementations of the Forward/Backward algorithms.

First, Forward/Backward values are calculated in an odds-ratio domain rather than the probability domain, so they are naturally pre-scaled to some extent. Each match emission probability $e_k(a)$ is replaced by its odds ratio $e_k(a)/f(a)$, the same ratio used for log-odds scores. By the same arguments used to analyze underflow of

the MSV implementation above, the odds ratio of the worst-case optimal local alignment path is a simple function of model and target lengths M and L, with a lower bound of about 2^{-60} (about 10^{-26}) if we assume $L,M \leq 10^6$ as design limits, well within the allowable range of an IEEE754 float representation.

Second, I exploit the fact that the HMMER3 profile HMM is a multihit local alignment model, not a glocal alignment model (glocal means global with respect to the model, local with respect to the target). When the profile HMM is a multihit local alignment model, a rescaling approach works. For any path with two aligned regions connected by a long deletion, there must exist an alternative path that counts the same two aligned regions as two local alignments, connected by a reinitiation path ($M_k \rightarrow E \rightarrow J \rightarrow B$). This reinitiation path, with only three transition probabilities on the same row, presents no underflow difficulties. The path with a long deletion may underflow, but then its complete path must be negligible relative to the alternative multihit local alignment path.

Determining the appropriate scale factor s_i requires examining each value in the row, which typically requires extra computation. I exploit the fact that the value in the cell for the E state has already calculated a maximum over all M_k states. It is sufficient to use the E cell value itself as the scale value s_i, setting the E cell value to 1 and rescaling all other values in the row.

Still, rescaling a row also requires extra computation. Here I exploit the fact that rescaling every row is unnecessary. Instead, when the E cell odds-ratio value exceeds a certain threshold, this triggers a rescaling event for that row. Other rows have $s_i = 1$.

I implemented Forward and Backward using sparse rescaling and striped SIMD vectors of four parallel 32-bit floats. Overall these implementations are about 16-fold faster than standard serial implementations using the log-sum lookup table operation. The overall 16-fold acceleration is likely a combination of about a 4-fold speedup from the SIMD vector parallelization with about 4-fold from replacing the log-sum operation with addition and multiplication. This makes the Forward/Backward algorithms only about 30-fold slower than the MSV filter.

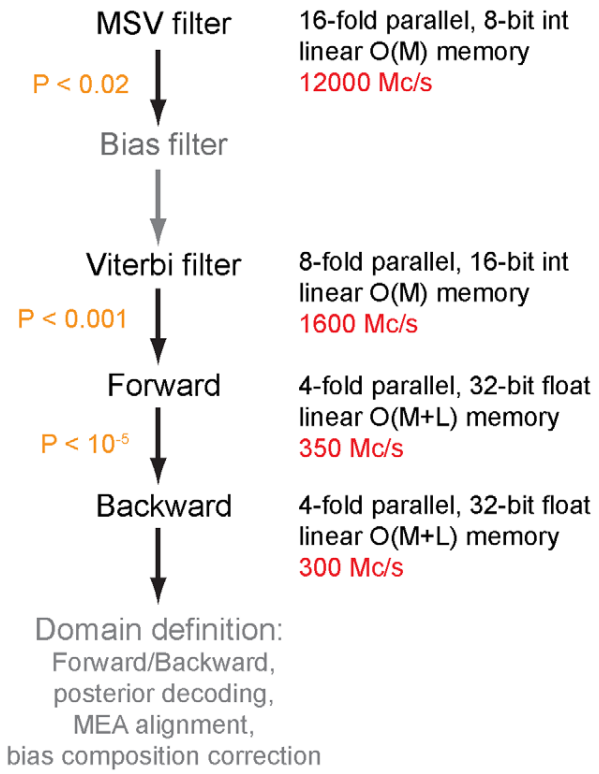


Figure 4. The HMMER3 acceleration pipeline. Representative calculation speeds are shown in red, in units of millions of dynamic programming cells per second (Mc/s). Default P-value thresholds for MSV, Viterbi, and Forward filtering steps are shown in orange. The bias filter and the domain definition steps are not described in detail in this manuscript, and are shown in gray.
doi:10.1371/journal.pcbi.1002195.g004

The HMMER3 acceleration pipeline

The MSV and Forward/Backward methods described above are implemented in the so-called "acceleration pipeline" at the core of the HMMER3 software implementation (http://hmmer.janelia.org). The acceleration pipeline is summarized in Figure 4. One call to the p7_Pipeline() function is executed for each model/sequence comparison.

The pipeline either accepts or rejects the entire comparison at each step, based on the P-value of the log-odds score. For example, by default the MSV filter passes if a comparison gets a P-value of less than 0.02 (i.e., the top-scoring 2% of random nonhomologous sequences are expected to pass the filter). I have not yet explored the more sophisticated approach of using alignment information from earlier and faster steps in the pipeline to constrain (band) subsequent dynamic programming calculations.

All P-value calculations assume that the query profile and target sequence have residue compositions close to the overall average for proteins. In some cases, a query profile has a biased composition, and this bias matches a bias found in many target database sequences. Membrane proteins, for example, are skewed towards a more hydrophobic composition, and tend to match other nonhomologous membrane proteins with scores higher than expected under a simple average-composition null hypothesis. HMMER3 has methodology for recalculating scores and P-values to compensate for biased composition, but this methodology (the so-called "null2" correction, not described here for reasons of space) is placed late in the pipeline because it is computationally intensive. At the MSV filter step, the uncorrected MSV P-value may be underestimated in biased composition matches, which means more than the expected fraction of nonhomologous sequences may passes the MSV filter, which in some cases can be sufficient to slow the pipeline. HMMER3 inserts a "bias filter" step to reduce this problem. The bias filter step is shown in gray in Figure 4 because it is not described in detail in this paper. Briefly, the bias filter calculates a fast, heuristic biased composition correction to the MSV filter score using a two-state hidden Markov model with one emission distribution set to the average protein residue composition and the other emission distribution set to the average composition of the query profile, fully connected by four arbitrary hand-tuned transition probabilities. The pipeline rescores the sequence with this correction applied, and retests the modified P-value against the MSV filter threshold of 0.02. The bias filter has no effect on the final reported score of a sequence, which is calculated by the full Forward algorithm; the bias filter only has the effect of making the MSV filter remove additional matches that appear to be due to biased composition.

To further reduce the computational load that arrives at the full Forward step, an additional filter, the Viterbi filter, was implemented and inserted in the pipeline. The Viterbi filter is a striped SIMD vector implementation of optimal gapped alignment to the profile (Figure 1A). It is implemented 8-fold vector parallel in 16-bit integers, because the numerical analysis of roundoff error accumulation is less favorable for Viterbi than for MSV and more precision is required. Following the same arguments as described

for MSV, Viterbi filter scores do not underflow (within design limits of $L,M \leq 10^6$) but may overflow the 16-bit representation, which is in units of 1/500 bits with an integer offset of 12,000, such that representable scores range from −89.5 to 41.5 bits. Any score that overflows the 41.5 bit upper limit is sure to pass any reasonable filter P-value anyway ($P(x>41.5)<10^{-12}$; the default Viterbi filter threshold is $P<0.001$).

The Forward and Backward algorithms for the pipeline are implemented in specialized efficient-memory forms called p7_ForwardParser() and p7_BackwardParser(). Each parser stores only a single row i of the dynamic programming matrix for each sequence residue x_i, plus the complete columns for the "special" states E, N, J, B, and C. This yields $O(M+L)$ linear-memory implementations, with sufficient stored information to allow Forward/Backward posterior decoding of the probable positions of B and E states on the target sequence, defining the probabilities of local alignment endpoints. A target sequence that passes the Viterbi filter is scored with the full Forward parsing algorithm. If the Forward score passes a P-value threshold (default $P<10^{-5}$), the Backward parser is calculated. Forward/Backward probabilities are used to estimate local alignment "regions" of substantial posterior probability mass in the target sequence. Each region is then subjected to a conceptually separate analysis pipeline, the "domain definition" pipeline, which identifies individual homologous regions and alignments, using a series of steps including full-matrix Forward/Backward, posterior decoding, maximum expected accuracy alignment, and a region-specific biased composition score correction. The domain definition procedure is not described in detail in this paper.

The acceleration pipeline is memory efficient. The MSV and Viterbi filters are only concerned with scores, not alignments, so they are implemented in linear-memory $O(M)$ forms that store only a single dynamic programming row. The Forward and Backward algorithms are used in a $O(M+L)$ "parser" form just described. The domain definition pipeline, however, is not memory efficient. It currently calculates full $O(ML')$ Forward/Backward and posterior decoding matrices for each identified subsequence (region) of length $L' \leq L$ in the target sequence. Until these steps in the domain postprocessor are replaced with more memory-efficient algorithms, HMMER3 can occasionally exhaust available memory on some large model/sequence comparisons.

Speed benchmarking

Figure 5 shows benchmark measurements of the speed of HMMER3, compared to the speed of BLAST [3], FASTA [40], SSEARCH (the FASTA implementation of Smith/Waterman), HMMER2, and the UCSC SAM profile HMM software [37]. Search speeds are shown in units of millions of dynamic programming cells calculated per second (Mc/s), measured on a single processor core (see Methods). The number of dynamic programming cells is the product of the query length M and the target database length L in residues. A straightforward implementation of dynamic programming sequence alignment scales in time as $O(ML)$, so reporting a speed in units of Mc/s is expected to be relatively independent of query and target length. In practice, the fastest search programs tend to show some additional dependence on query sequence length, with more efficient performance on longer queries. Figure 5 looks at a range of different queries of different lengths.

To measure the "typical" performance of each program, without complicating variation arising from producing the voluminous alignment output for some queries that hit large protein superfamilies, panel A (left) shows benchmarks done on random (shuffled) target sequences. The panel shows results for 76 query profiles (or representative single sequences), chosen to sample the full range of query lengths in the Pfam protein domain database from 7 to 2,217 residues. These results show that HMMER3 performance is comparable to other fast database search programs; somewhat slower (by about 2- to 3-fold) than NCBI BLAST, and somewhat faster (by about 3-fold) than WU-BLAST, for example. The speed of SSEARCH, the Smith/Waterman local alignment implementation in the FASTA package, is worth noting in this figure; SSEARCH has recently been accelerated by implementing Farrar's striped SIMD vector

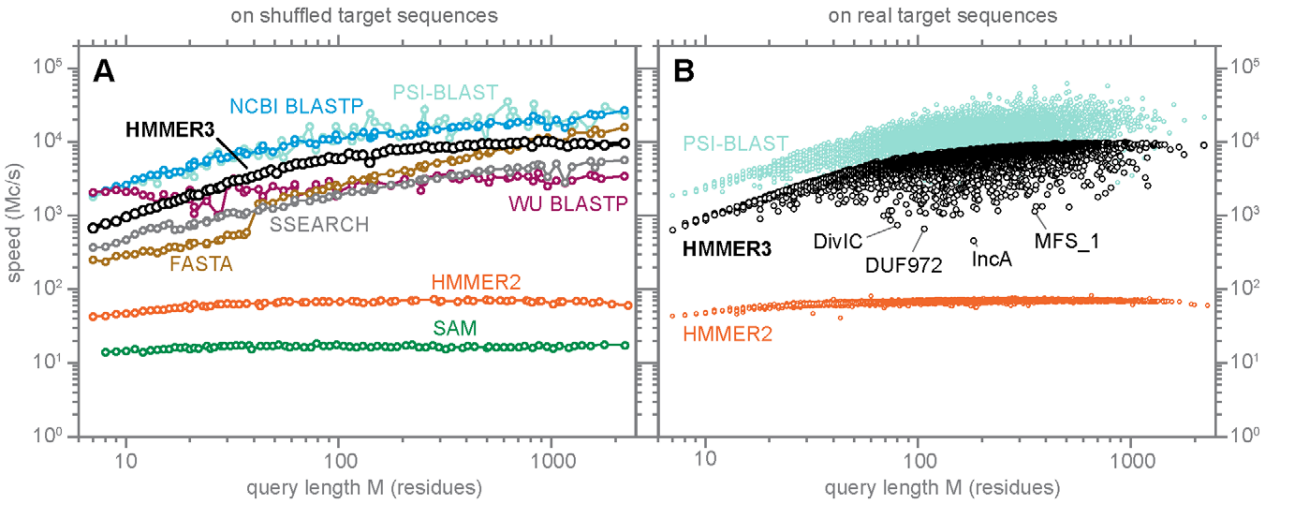


Figure 5. Speed benchmarks. Each point represents a speed measurement for one search with one query against 10^6 target sequences (10^4 for the slow HMMER2 and SAM programs, 10^5 for FASTA and SSEARCH), on a single CPU core (see Methods for more details). Both axes are logarithmic, for speed in millions of dynamic programming cells per second (Mc/s) on the y-axis and query length in residues on the x-axis. Panel A shows "typical best performance" speed measurements for several different programs including HMMER3, for 76 queries of varying consensus lengths, chosen from Pfam 24, for searches of randomized (shuffled) target sequences. Panel B shows a wider range of more realistic speed measurements for all 11,912 profiles in Pfam 24, on searches of real target protein sequences from UniProt TrEMBL.
doi:10.1371/journal.pcbi.1002195.g005

methods, allowing it to achieve speed comparable to the heuristic FASTA and WU-BLAST programs. HMMER3 is faster than HMMER2 by up to 140×, even though HMMER3 calculates full Forward scores whereas HMMER2 calculated faster Viterbi optimal alignment scores. Compared to SAM, which does calculate full Forward scores, HMMER3 is about 600× faster.

To measure a wider and more realistic range of real-world performance, panel B (right) shows benchmarks for 11,912 different queries (every Pfam 24 profile) on real sequences from UniProt TrEMBL. Programs that simply do a dynamic programming alignment to each target sequence, such as HMMER2 (orange points) or SSEARCH (not shown), show performance essentially independent of the properties of the query and target sequences. Programs that use heuristics and filters, however, are sensitive to how well a given search obeys the assumptions of the heuristic and/or filter thresholds. Both HMMER3 and PSI-BLAST speed vary not only by query length, but also vary substantially around their average for a given query length. PSI-BLAST speed in panel B varies both up (by up to about 3-fold) and down (by up to about 10-fold) from its "typical" performance in panel A, presumably reflecting variation in how many word hits and hit extensions need to be processed for a given search. HMMER3 speed tends to vary only downwards from its typical performance, by up to about 20-fold. In panel B, I highlight examples of four poorest-performing HMMER3 searches, on the DivIC, DUF972, IncA, and MFS_1 models. Even with the bias filter step included in the acceleration pipeline, the dominant cause of poorer HMMER3 search performance remains biased composition sequences (such as transmembrane proteins) in which more comparisons pass the fast filters of the acceleration pipeline than expected by P-value calculations that assume average target sequence compositions, causing more comparisons to reach the compute-intensive Forward/Backward calculations.

Sensitivity/specificity benchmarking

A filtering approach will generally compromise search sensitivity by some degree, because a filter will erroneously remove true homologs at some rate. We want this rate to be negligible. To measure how much search sensitivity is attenuated by the use of the MSV filter and the HMMER3 acceleration pipeline, I performed benchmarks to compare sensitivity/specificity of default HMMER3 hmmsearch (with the acceleration filter pipeline) to hmmsearch --max, an option that turns off all the filters and runs the full Forward scoring algorithm on every target sequence. I also benchmarked HMMER2 and several other homology search programs for comparison. These results are shown in Figure 6.

These benchmarks are automatically and semi-randomly generated by a program (create-profmark). The program starts from a source of trusted alignments of homologous protein domains (here, Pfam 24 seed alignments), a source of typical full-length protein sequences (here, UniProt SwissProt 2011_03), and a choice of method for synthesizing nonhomologous sequence segments (such as shuffling a randomly chosen segment of a UniProt sequence). A query alignment and a set of true test domains (trusted to be homologous to the query) is created by applying single-linkage-clustering by percent identity to a Pfam alignment, and using that clustering to select sequences such that no true test domain has more than 25% pairwise identity to any sequence in the query alignment, and no more than 50% pairwise identity to any other test domain. True test sequences are created by concatenating one or two test domains together with nonhomologous sequence segments, with a total sequence length sampled from the distribution of UniProt sequences. False (nonhomologous) test sequences are created by concatenating nonhomologous sequence segments, with both segment length and total sequence length sampled from the length distributions of the true test sequences. The procedure, its rationale, and some of its caveats are described in more detail in Methods.

To benchmark a profile method (HMMER, SAM, or PSI-BLAST), a profile is built from each query alignment and searched against the target database of test sequences and decoys. The results of all searches (for all different queries) are merged and sorted by E-value, and this ranked list is used to calculate a plot of fraction of true positives detected at increasing thresholds of false positives per query from 0.001 to 10.

To benchmark a single-sequence query method (BLAST, SSEARCH, FASTA) more fairly against profile methods, a family-pairwise-search (FPS) method is used [41] (as opposed to selecting just one query sequence from the alignment). Each individual sequence in the query alignment is searched against the target database; the best E-value found to any query sequence is treated as the E-value of the target sequence, and results for all queries are merged, sorted, and treated as above.

Figure 6 shows results for two different benchmarks. Panel A shows results where true test sequences have a single embedded domain, and nonhomologous sequences are synthesized as i.i.d. (independent identically distributed) random sequence from the average UniProt residue frequency distribution; there are 2,141 query alignments, 11,547 true test sequences, and 200,000 decoys. Because nonhomologous sequences are unrealistically simple (no biased composition, no repetitive sequence), this benchmark does not exercise the various corrections for biased composition that some programs have (such as HMMER and BLAST), and is more of a baseline test of the best-case sensitivity of a search program. Panel B shows results of my default profmark benchmark, where there are two embedded domains per true test sequence (in order to test the ability of a program to correctly detect and align multiple domains per sequence, although such results are not shown here), and where nonhomologous sequence segments are created by shuffling randomly chosen segments of UniProt sequences. Because nonhomologous segments in this benchmark can show more realistic monoresidue composition biases (though not higher-order bias such as tandem repeats), this version of the benchmark is more realistic and exercises more *ad hoc* parts of programs that try to correct for biased composition. Panel A shows results that are largely independent of biased composition issues, but less realistic. Panel B shows results that are more dependent on the ability of a program to handle biased composition, and probably more realistic.

The main result in both panels is that sensitivity and specificity are essentially identical for HMMER3 either with the acceleration pipeline (dark black lines) and without it (--max; red lines). There is a slight loss in sensitivity caused by the acceleration pipeline, but this loss is more than compensated by the gain in sensitivity of HMMER3 over HMMER2 (either in its default "glocal" alignment search mode or its local alignment search mode; orange circles and squares respectively). At high specificity (low false positive rates) on more realistic biased decoys, default HMMER3 can appear to be better than unfiltered HMMER3 (hmmsearch --max) because the bias filter removes some problematic biased false positive decoys that the supposedly more powerful biased composition corrections in HMMER3 Forward scores fail to correct.

Benchmarks for other programs (such as BLAST, SAM, FASTA, and SSEARCH) are shown only for the sake of rough comparison. The intent is not to thoroughly benchmark HMMER against these programs, but to provide an additional sense of scale, putting the difference between HMMER3 with and without its acceleration pipeline in context – that is, showing that the

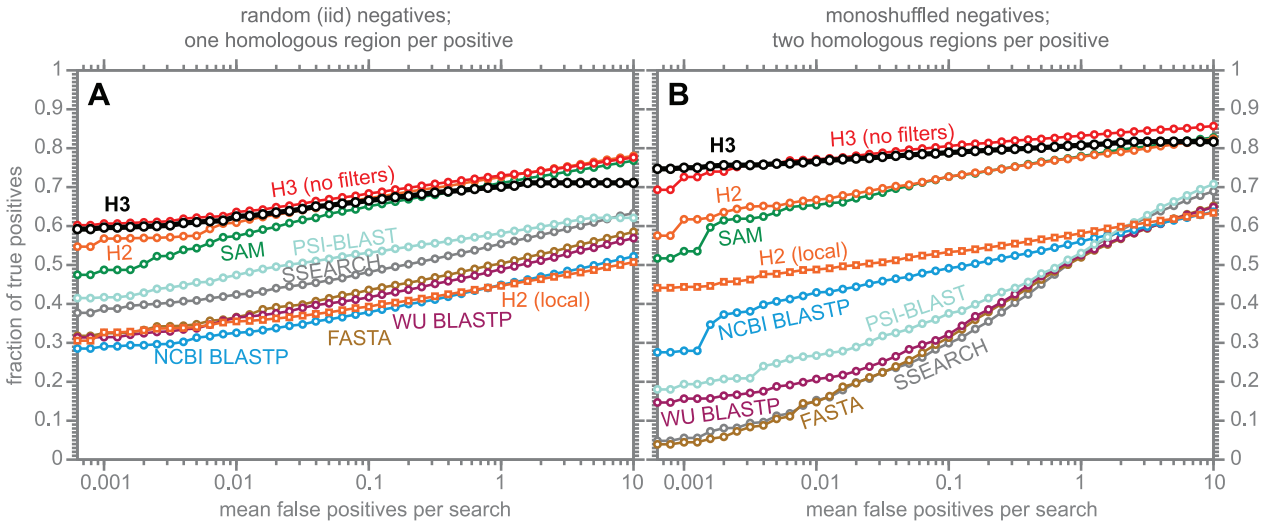


Figure 6. Benchmark of search sensitivity and specificity. For different programs, searches are performed either by constructing a single profile from the query alignment (HMMER3, HMMER2, SAM, PSI-BLAST), or by using "family pairwise search" [41] in which each individual sequence is used as a query and the best E-value per target sequence is recorded (BLASTP, SSEARCH, FASTA). In each benchmark, true positive subsequences have been selected to be no more than 25% identical to any sequence in the query alignment (see Methods). Panel A shows results where nonhomologous sequence has been synthesized by a simple random model, and each true positive sequence contains a single embedded homologous subsequence (a total of 2,141 query multiple alignments, 11,547 true positive sequences, and 200,000 decoys). Panel B shows results where nonhomologous sequence is synthesized by shuffling randomly chosen subsequences from UniProt, and each true positive contains two embedded homologous subsequences (a total of 2,141 query alignments, 24,040 true positive sequences, and 200,000 decoys). The Y-axis is the fraction of true positives detected with an E-value better than the number of false positives per query specified on the X-axis.
doi:10.1371/journal.pcbi.1002195.g006

difference between HMMER3 with and without its acceleration pipeline is minor, compared to differences among programs. These programs were only run in their default configuration. I did not explore available options that might improve their performance on this benchmark. Although I believe the results to be fair and representative, I have to interpret these results with caution. I benchmark HMMER routinely on these benchmarks during development. It is impossible to avoid some degree of training on the benchmark, even though the benchmarks are somewhat randomized. Nonetheless, some informative trends in these results agree with previous independent published benchmarking from other groups, so the comparisons are probably useful as a rough guide. For example, Madera and Gough published a benchmark [28] in which they concluded that SAM significantly outperformed HMMER2, which at the time disagreed with my experience. Johnson [29] traced this to the fact that Madera and Gough had switched HMMER2 from its default glocal alignment mode to its nondefault local search mode, which we had not spent much time testing or tuning, and we had not previously realized how sensitive local alignment is to the model's information content per position. This led us to realize how important SAM's entropy-weighting technique is for local alignment, whereas it is much less important in glocal alignment [29]. This story is reflected in the benchmarks in Figure 6, where HMMER2 *local* alignment performs poorly relative to SAM, HMMER2 *glocal* alignment is comparable to SAM, and HMMER3 local alignment (with the entropy-weighting technique) is perhaps a bit better than SAM. Most of the difference between HMMER3 and SAM is in the high-specificity regime of the more realistic benchmark that includes biased-composition segments (Panel B), and thus is likely to result from differences in the *ad hoc* bias composition corrections that differ between SAM and HMMER, rather than any fundamental difference in their profile HMM parameterizations or their Forward scoring algorithms, which I believe are quite similar.

Another story reflected in these benchmarks is about the widespread belief that full Smith/Waterman alignment is superior to BLAST's fast heuristic approximation of Smith/Waterman. This is true in the easier benchmark (Panel A) but not when decoy sequences include biased composition segments (Panel B). Indeed, in Panel A all three fast sequence search programs (WU-BLAST, NCBI-BLAST, and FASTA) perform comparably (and worse than SSEARCH), whereas in Panel B, NCBI BLAST outperforms WU-BLAST, FASTA, and SSEARCH. This again seems likely to be showing the importance of biased composition score corrections. Biased composition correction has received close attention in NCBI BLAST software development [7,8,42], but is not part of the textbook description of "optimal" Smith/Waterman local alignment.

My main conclusion (that the acceleration pipeline has a negligible impact on the sensitivity/specificity of HMMER3 compared to unaccelerated Forward scoring) is supported by a more direct experiment. I searched all 11,912 Pfam 24 profile HMMs against the 516,081 sequences in UniProt SwissProt 2011_03 using five different option settings to hmmsearch, starting with --max and then successively turning on one filter step at a time in the acceleration pipeline (MSV, bias, Viterbi, and Forward), up to the default configuration with all four filter steps on. With --max, a total of 799,893 hits were found with an E-value of 0.0001 or less. Turning on the MSV filter loses 718 hits (0.09%). Overall, the default pipeline with all filter steps loses 2,701 hits (0.3%). Differences in significant hits are not necessarily all due to true homologs. It is possible for the unfiltered search to find a false positive that one or more of the filters would remove. However, the majority of these differences appear to be true homologs that are removed by the filters. Other than the 718 hits removed by the MSV filter, the great majority of the other losses are due to the bias filter inappropriately removing sequences that have strong biased compositions, but also contain a true homology region. The

most obvious problems with HMMER3 sensitivity/specificity seem to lie in its bias filter and bias composition score corrections, rather than in the use of the MSV filter as its primary acceleration.

Discussion

In describing the MSV heuristic and other acceleration methods implemented in HMMER3, I have not addressed the question of whether the MSV heuristic is better or worse than other heuristics, such as those in BLAST or FASTA. In sensitivity/specificity benchmarks (Figure 6), BLAST and FASTA perform about the same as unaccelerated Smith/Waterman, and HMMER3 performs about the same with and without its acceleration pipeline. This show that the overall sensitivity and specificity of these programs are not limited by their respective heuristics, but rather by their fundamental (unaccelerated) sequence comparison methods. Thus a better heuristic would be unlikely to improve overall sensitivity/specificity. A better heuristic would be a faster one with the same sensitivity/specificity. In a question about speed, we cannot satisfactorily show that one algorithm is necessarily faster than another. We can only rigorously compare particular implementations. Trying to mix and match different heuristic approaches and make definitive comparisons of their fully optimized speeds requires a time-consuming engineering and optimization effort dedicated to each implementation. So it remains an open question whether, for example, BLAST-style heuristics could be tuned to have enough sensitivity/specificity to match HMMER3 performance while still being faster. I developed MSV because it was easy for me, not because I tried different heuristics and found MSV to be better. I was working on striped SIMD vectorization of the Viterbi algorithm, and an MSV implementation is easily derived from striped SIMD Viterbi just by deleting the code that handles deletions and insertions. The heuristic approaches in BLAST and FASTA have the advantage of focusing subsequent slower computations on particular diagonals, whereas in HMMER's current approach, we wastefully recalculate full sequence alignments at each step of the acceleration pipeline. I expect that it will be fruitful to develop heuristics focused around high-likelihood diagonals, as BLAST and FASTA do, while using HMMER's SIMD vectorization methods.

Although this paper is about the acceleration methods used in HMMER3, HMMER3 also appears to be more sensitive than HMMER2. The main reason for this is the adoption of "entropy-weighting", a method introduced by the UC Santa Cruz group in the SAM profile HMM package [29,37], where the information content per position is reduced to a specified target number of bits. A second reason is the switch from Viterbi optimal alignment scores to Forward scores summed over the alignment ensemble [29].

On the other hand, I believe that the switch from default glocal alignment in HMMER2 to local alignment in HMMER3 has probably compromised some search sensitivity ("glocal" means global in the query, local in the target sequence: requiring a full-length domain alignment). Restoring glocal alignment to HMMER3 should improve search performance for profiles that are expected to match over their entire length, such as Pfam protein domain models. However, the fast E-value statistics for Forward and Viterbi scores (including MSV filter scores) are only valid for local alignment, and the numeric underflow analysis of the sparse rescaling technique in the Forward/Backward implementation assumes local alignment. Both problems will need to be addressed before glocal alignment is implemented usefully in HMMER3.

Here I have only described single-core performance. I have not discussed parallelization across multiple cores. HMMER3 search programs include rudimentary implementations of POSIX threads and MPI parallelization (message-passing in a cluster of computing nodes). These implementations currently scale poorly, to only modest numbers of processor cores (2–4 for multithreading, for example). Improved parallelization is a priority for future development.

HMMER3's handling of biased composition sequences is problematic. I chose to introduce an *ad hoc* "bias filter" into the acceleration pipeline, to deal with a small number of profiles that let too many sequences through the MSV filter and bog down in the slow Forward/Backward stages of the pipeline. The bias filter occasionally filters true positive hits. A disturbing failure mode can occur when a target sequence consists of a homologous subsequence surrounded by a large amount of nonhomologous biased composition sequence; in this case, the bias filter may aggressively remove the entire sequence. Although other database search programs have analogous issues with over-aggressive composition masking, one future focus for HMMER3 development will be on improving its formal probability model of *non*homologous sequence.

This paper describes an initial baseline for HMMER3 speed performance on a single processor core. The prospects for substantial future improvements are good. There are many obvious opportunities for incremental optimizations. Bjarne Knudsen (CLCbio, Aarhus, Denmark) has already contributed an important optimization of the MSV filter that increases overall HMMER3 speed by about two-fold. The Knudsen optimization will appear in the next HMMER code release, and we will likely describe it in a future manuscript. Another optimization opportunity is to preprocess the target database sequence file into an efficient binary format, as BLAST does with its BLAST databases. HMMER3 still reads standard flatfile sequence databases, such as FASTA and UniProt text formats. Another optimization opportunity is to convert the filters in the pipeline from their current mode of filtering entire target sequences (which was easy to implement) to instead store and retrieve more information about the location of alignment probability mass, so subsequent steps (including Forward/Backward) can be done as banded dynamic programming calculations within high-probability envelopes, as opposed to reprocessing the entire query/target comparison at each pipeline step. Because of these and other straightforward optimization opportunities, I expect HMMER3 speed will surpass NCBI BLAST speed in the relatively near future.

This speed makes it feasible to apply profile HMM technology to standard sequence database searches with single sequence queries, including iterative database searches. A position-independent scoring system for single sequences is just a special case of a profiled position-specific scoring system. A "profile" HMM can be built from a single sequence, using position-independent probabilities obtained from standard scoring matrices such as BLOSUM62, plus a couple of parameters for gap-open and gap-extend probabilities. The HMMER3 software package includes a program *phmmer* for protein database searches akin to *blastp*, and a program *jackhmmer* for iterative protein database searches akin to *psiblast*. These programs, their parameterization, and the effect of extending profile HMM technology, the Forward algorithm, and probabilistic inference methods to routine sequence database searches will be described elsewhere.

Materials and Methods

Software implementation and availability

HMMER3 is implemented in POSIX ANSI/ISO C99. Vector implementations are provided for Intel-compatible (SSE) and

PowerPC (Altivec/VMX) processors. A "dummy" non-vectorized implementation is provided for other processors, sufficient to enable compilation but about 100× slower than normal HMMER3 vectorized performance. HMMER3 is portable to any POSIX-compliant operating system, including Linux and Mac OS/X, and also to Windows if an optional POSIX compatibility package has been installed (such as Cygwin). It builds from source using a standard GNU configure script and UNIX make. It includes a suite of automated tests written in Perl, C, and Bourne shell. User documentation is provided as PDF and man pages. Source code is freely available under a GNU GPLv3 license. Precompiled binary distributions are available for several platforms including Intel/Linux, Intel/MacOSX, and Intel/Windows. A web interface to HMMER3-based database searches is available, including batch searches and RESTful web services, hosted on computational resources supported by the Howard Hughes Medical Institute. All these resources are at http://hmmer.janelia.org.

Software and database versions used

Software versions used: SAM 3.5 (Jul 2005) [37], NCBI BLAST+ 2.2.24+ (Aug 2010) [3], FASTA 36.3.3 (Feb 2011) [40], WU-BLAST 2.0MP-WashU (May 2006), HMMER 2.3.2 (Oct 2003), and HMMER 3.0 (Mar 2010).

Example sequence alignments and profile HMMs were sampled from Seed alignments and profiles in Pfam 24 [11]. Example target sequences were sampled from UniProt version 2011_03 [43]. One experiment that characterized roundoff error used older versions, Pfam 22 and UniProt 7.0.

Speed benchmarking

All program timings were measured in total (wall clock) time on a single execution thread (single core) of a dedicated and unloaded cluster node, where a node has eight 2.66 GHz Intel Gainestown X5550 cores and 24 GB RAM. The same search was run twice sequentially and timed in the second run, to allow filesystem caching of target databases.

For speed benchmarks of programs that take a single sequence as a query (instead of an alignment), the median length sequence was extracted from the query alignment.

Construction of the sensitivity/specificity benchmark

Sensitivity/specificity benchmarks were created with the create-profmark program, included in the HMMER3 source code. This program allows construction of the wide range of different and randomized benchmarks used during HMMER development. My concern is that because benchmarking is repeated at every step of code development, it is nearly impossible for a developer to avoid overtraining on any in-house benchmark. Synthesizing a variety of partially randomized benchmarks helps mitigate this effect somewhat, compared to relying on a single static benchmark.

The profmark benchmarks use a set of trusted multiple alignments (such as Pfam Seed alignments) as a source of both *query multiple alignments* and distantly related *true test sequences* trusted to be homologous. The individual sequences in each input alignment are clustered by pairwise percent identity, and different clusters are selected to be queries versus test sequences such that no true test domain has more than 25% identity to any sequence in the query alignment, and no true test domain has more than 50% identity to another true test domain. To create realistic-length true test sequences, and to challenge the ability of a program to detect homologous local alignments in a larger target sequence, true test sequences are synthesized by embedding one or two test domains in a larger nonhomologous sequence.

Using a sequence database like Pfam instead of a 3D structure database like SCOP or CATH as a source of trusted true homology relationships has the advantage that a more challenging variety of sequences is tested. Structure databases are biased toward well-ordered globular domains. A weakness of a sequence-based benchmark is that "true homologs" are inferred by current computational sequence comparison methods, rather than being defined by an independent criterion like 3D structure comparison. In particular, my profmark benchmarks are constructed from Pfam alignments as a trusted definition of true homologs, and Pfam is itself constructed with HMMER. There may be some danger that this circularity creates a bias against other search programs. Specifically the danger is that if there are remote homologs that are undetected by profile HMM methods, but that could be detected by another method, any such sequences have been selected against in Pfam. Intuitively, I think this danger is negligible. If a search program was sufficiently powerful that it could detect homologs excluded from Pfam, it ought to be even better at detecting the closer homologs that were included and artificially separated (by profmark's clustering procedure) into challengingly dissimilar query alignment and true test domains. Empirically too, any danger has seemed negligible, because profmark benchmarks tend to be broadly concordant with other published benchmarks [28,44]. Nonetheless, I have more confidence in using profmark benchmarks for internal comparisons (HMMER vs. HMMER, for different option settings) than for comparisons to other search programs.

False (decoy) sequences (including the nonhomologous flanking sequence around embedded test domains in true test sequences) are created synthetically. If we are trying to find methods that detect previously undetectable homologies, no source of real biological sequences will ever be reliably known to be nonhomologous to the benchmark, and we certainly do not want to penalize a powerful method that identifies new true relationships that are currently annotated as nonhomologous "false positives" [29]. One disadvantage of synthetic nonhomologous sequence is that it is difficult to create realistic sequences with the same challenging properties of real biological sequences, such as biased composition and repetitive sequence.

In detail, the profmark creation procedure is the following, starting from a source of multiple alignments (usually Pfam seeds) and a source of typical single target sequences (usually UniProt/SwissProt):

- Convert all degenerate residue characters to X. (Although HMMER reads all standard degeneracy codes for protein and nucleic acid sequences, some search programs do not.)
- Remove sequence fragments. By default, any sequence of length less than 70% of the mean unaligned sequence length in the alignment is defined as a fragment.
- Cluster the sequences by single-linkage clustering at a default threshold of $\geq 25\%$ percent identity (defined as the number of identical residues divided by the shorter sequence length, in the given pairwise alignment). Between any two clusters, there is no pair of sequences closer than 25% identity. If there is only one cluster, exclude this alignment from the benchmark and skip to the next alignment. Define the largest cluster as the *query*. Save it to a file, in its original multiple alignment.
- Cluster the remaining sequences by single-linkage clustering at a default threshold of $\geq 50\%$ identity. If there are less than two clusters, exclude this alignment from the benchmark and skip to the next alignment. From each cluster, select one sequence at random. These are the *true test domains*.

- Create synthetic true test sequences by embedding one or two true test domains in a larger nonhomologous sequence. The total length of the test sequence is sampled from the length distribution of the input sequence database, conditional on being at least as long as the true test domain(s). True test domains are inserted at randomly sampled locations in the sequence. The remaining two or three nonhomologous sequence segments are synthesized as described below. Thus true test sequences are composed either of three segments (one homologous, two not) or five segments (two homologous, three not).
- The program implements a choice of several different methods for generating nonhomologous sequence segments. The default is "monoshuffling": to select a sequence segment at random from the input sequence database and shuffle its residues, preserving 0th-order residue composition and bias. Figure 6 also shows the use of i.i.d. (independent identically distributed) synthetic sequence with each residue simply sampled from the average residue frequency distribution of proteins. Other options include reversed sequences and shuffling while preserving di-residue composition. Though more realistic, and useful when looking carefully and manually for failure modes, di-residue shuffling and reversed sequences are problematic as a source of nonhomologous segments in automated benchmarking. Exact di-residue shuffling preserves significant sequence identity to the original sequence over surprising segment lengths, and reversed sequences are surprisingly significantly more likely to show a significant match to the original sequence (because of a counterintuitive statistical effect of the frequency of approximate palindromes in any sequence).
- Decoys (negative sequences) are created by randomly selecting a true test sequence (solely to obtain its three or five segment lengths – not its sequence) then concatenating nonhomologous segments of the same lengths. The length distribution of negative sequences, and the length distribution of potentially biased nonhomologous subsequences embedded in them, is therefore matched to the distributions for the true test sequences.

For the experiments in Figure 6, the create-profmark procedure was applied to 11,912 Pfam 24 seed alignments [11] and the UniProt/SwissProt sequence database (version 2011_03, 516,081 sequences) [43] either with options –iid –mono (Figure 6A) or default (Figure 6B). The benchmark in panel A is composed of 2,141 query alignments, 24,040 true test sequences containing single homologous domains, and 200,000 decoys. The benchmark in panel B is composed of 2,141 query alignments, 11,547 true test sequences containing two homologous domains, and 200,000 decoys.

Acknowledgments

Jeremy Buhler and Christopher Swope (Washington University, St. Louis) contributed important optimizations for the Viterbi and Forward implementations in HMMER2. Erik Lindahl (Stanford) contributed the first SIMD vector implementation of HMMER2, for PowerPC processors with VMX/Altivec instructions. Alex Coventry (Cornell), Steve Johnson (Washington Univ., St. Louis), and Elon Portugaly (Hebrew University) did exploratory work that led to the HMMER3 project. Bjarne Knudsen (CLCbio, Denmark) has contributed several optimizations and suggestions based on a careful analysis of the HMMER3 source code. Elena Rivas (Janelia Farm) contributed numerous clarifications and improvements to the source code in the course of reimplementing the core of HMMER3 for continuous emission variables for modeling mouse vocalizations. Although space permits only these brief personal acknowledgments, a great many others have contributed to the development of HMMER3, and I also thank this wide community of developers and users. Goran Ceric provided masterful support of Janelia Farm's high performance computing resources. Heidi Henning, HHMI deputy general counsel, provided invaluable legal assistance on open source issues. HMMER is a trademarked open source project supported by the Howard Hughes Medical Institute. I dedicate this paper to the memory of my colleague and friend Michael Farrar, principal software engineer on the HMMER3 project, who died unexpectedly in December 2010.

Author Contributions

Conceived and designed the experiments: SRE. Performed the experiments: SRE. Analyzed the data: SRE. Contributed reagents/materials/analysis tools: SRE. Wrote the paper: SRE.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSIBLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
3. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10: 421.
4. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology: Applications to protein modeling. J Mol Biol 235: 1501–1531.
5. Durbin R, Eddy SR, Krogh A, Mitchison GJ (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge UK: Cambridge University Press.
6. Altschul SF, Bundschuh R, Olsen R, Hwa T (2001) The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res 29: 351–361.
7. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994–3005.
8. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, et al. (2005) Protein database searches using compositionally adjusted substitution matrices. FEBS J 272: 5101–5109.
9. Yu YK, Gertz EM, Agarwala R, Schäffer AA, Altschul SF (2006) Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. Nucleic Acids Res 34: 5966–5973.
10. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211–D215.
11. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.
12. Chaudhary V, Liu F, Matta V, Meng X, Nambiar A, et al. (2006) Parallel implementations of local sequence alignment: Hardware and software. In: Zomaya AY, ed. Parallel Computing for Bioinformatics and Computational Biology. Sydney, Australia: John Wiley & Sons, Inc., chapter 10. pp 233–264.
13. Walters JP, Qudah B, Chaudhary V (2006) Accelerating the HMMER sequence analysis suite using conventional processors. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications; 18–20 April 2006; Vienna, Austria. AINA 2006. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/AINA.2006.68. Accessed 22 August 2011.
14. Landman J, Ray J, Walters JP (2006) Accelerating HMMer searches on Opteron processors with minimally invasive recoding. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications; 18–20 April 2006; Vienna, Austria. AINA 2006. Available: http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber = 1620450. Accessed 22 August 2011.
15. Sachdeva V, Kistler M, Speight E, Tzeng THK (2007) Exploring the viability of the Cell Broadband Engine for bioinformatics applications. Parallel Comput 34: 616–626.
16. Maddimsetty RP (2006) Acceleration of Profile-HMM Search for Protein Sequences in Reconfigurable Hardware [Master's thesis]. (St. Louis): Department of Computer Science and Engineering, Washington University in St. Louis. 70 p.
17. Derrien S, Quinton P (2007) Parallelizing HMMER for hardware acceleration on FPGAs. In: Proceedings of the 18th IEEE International Conference on Application-Specific Systems, Architectures, and Processors; 9–11 July 2008;

Montreal, Quebec. ASAP 2007. Available: http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber = 4429951. Accessed 22 August 2011.
18. Jacob A, Lancaster J, Buhler J, Chamberlain RD (2007) Preliminary results in accelerating profile HMM search on FPGAs. In: Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium; 26–30 March 2007; Long Beach, California, United States. IPDPS 2007. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/IPDPS.2007.370447. Accessed 22 August 2011.
19. Oliver T, Yeow LY, Schmidt B (2007) High performance database searching with HMMer on FPGAs. In: Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium; 26–30 March 2007; Long Beach, California, United States. IPDPS 2007. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/IPDPS.2007.370448. Accessed 22 August 2011.
20. Horn DR, Houston M, Hanrahan P (2005) ClawHMMER: A streaming HMMer-search implementation. In: Proceedings of the ACM/IEEE Conference on Supercomputing; 12–18 November 2005; Seattle, Washington, United States. SC 2005. Available: http://dl.acm.org/citation.cfm?id = 1105776. Accessed 22 August 2011.
21. Walters JP, Balu V, Kompalli S, Chaudhary V (2009) Evaluating the use of GPUs in liver image segmentation and HMMER database searches. In: Proceedings of the 23rd IEEE International Symp. on Parallel & Distributed Processing; 23–29 May 2009; Rome, Italy. IPDPS 2009. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/IPDPS.2009.5161073. Accessed 22 August 2011.
22. Chukkapalli G, Guda C, Subramaniam S (2004) SledgeHMMER: A web server for batch searching the Pfam database. Nucleic Acids Res 32: W542–W544.
23. Rekapalli B, Halloy C, Zhulin IB (2009) HSP-HMMER: A tool for protein identification on a large scale. In: Proceedings of the 24th ACM Symposium on Applied Computing; 9–12 March 2009; Honolulu, Hawaii, United States. SAC 2009. Available: http://dl.acm.org/citation.cfm?id = 1529282.1529443. Accessed 22 August 2011.
24. Sun Y, Buhler J (2007) Designing patterns for profile HMM search. Bioinformatics 23: e36–e43.
25. Sun Y, Buhler J (2009) Designing patterns and profiles for faster HMM search. IEEE/ACM Trans Comput Biol Bioinform 6: 232–243.
26. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11: 431.
27. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197.
28. Madera M, Gough J (2002) A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 30: 4321–4328.
29. Johnson S (2006) Remote Protein Homology Detection Using Hidden Markov Models [Ph.D. thesis]. (St. Louis): Washington University School of Medicine, Washington University in St. Louis. 92 p.
30. Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17: 117–125.
31. Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4: e1000069.
32. Rognes T (2001) ParAlign: A parallel sequence alignment algorithm for rapid and sensitive database searches. Nucleic Acids Res 29: 1647–1652.
33. Farrar M (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23: 156–161.
34. Wozniak A (1997) Using video-oriented instructions to speed up sequence comparison. Comput Applic Biosci 13: 145–150.
35. Rognes T, Seeberg E (2000) Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. Bioinformatics 16: 699–706.
36. Milosavljević A, Jurka J (1993) Discovering simple DNA sequences by the algorithmic significance method. Comput Applic Biosci 9: 407–411.
37. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14: 846–856.
38. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–286.
39. Melnikoff SJ, Quigley SF (2003) Implementing the log-add algorithm in hardware. Electronics Letters 12: 939–941.
40. Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Meth Mol Biol 132: 185–219.
41. Grundy WN (1998) Homology detection via family pairwise search. J Comput Biol 5: 479–491.
42. Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF (2006) Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. BMC Biol 4: 41.
43. The UniProt Consortium (2011) Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39: D214–D219.
44. Price GA, Crooks GE, Green RE, Brenner SE (2005) Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. Bioinformatics 21: 3824–3831.

nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-023-01773-0

Fast and accurate protein structure search with Foldseek

Received: 17 February 2022

Accepted: 30 March 2023

Published online: 8 May 2023

Check for updates

Michel van Kempen[1,6], Stephanie S. Kim[2,6], Charlotte Tumescheit[2], Milot Mirdita[1,2], Jeongjae Lee[2], Cameron L. M. Gilchrist[2], Johannes Söding[1,3] ✉ & Martin Steinegger[2,4,5] ✉

As structure prediction methods are generating millions of publicly available protein structures, searching these databases is becoming a bottleneck. Foldseek aligns the structure of a query protein against a database by describing tertiary amino acid interactions within proteins as sequences over a structural alphabet. Foldseek decreases computation times by four to five orders of magnitude with 86%, 88% and 133% of the sensitivities of Dali, TM-align and CE, respectively.

The recent developments in in silico protein structure prediction at near-experimental quality[1,2] are advancing structural biology and bioinformatics. The European Bioinformatics Institute already holds over 214 million structures predicted by AlphaFold2 (ref. 3), and the ESM Atlas contains over 617 million metagenomic structures predicted by ESMFold[4]. The scale of these databases poses challenges to state-of-the-art analysis methods.

The most widely used approach to protein annotation and analysis is based on sequence similarity search[5–8]. The goal is to find homologous sequences from which properties of the query sequence can be inferred, such as molecular and cellular functions and structure. Despite the success of sequence-based homology inference, many proteins cannot be annotated because detecting distant evolutionary relationships from sequences alone remains challenging[9].

Detecting similarity between protein structures by three-dimensional (3D) superposition offers higher sensitivity for identifying homologous proteins[10]. The availability of high-quality structures for any protein of interest allows us to use structure comparison to improve homology inference and structural, functional and evolutionary analyses. However, despite decades of effort to improve speed and sensitivity of structural aligners, current tools are much too slow to cope with today's scale of structure databases.

Searching with a single query structure through a database with 100 million protein structures would take the popular TM-align[11] tool a month on one CPU core, and an all-versus-all comparison would take 10 millennia on a 1,000-core cluster. Sequence searching is four to five orders of magnitude faster: an all-versus-all comparison of 100 million sequences would take MMseqs2 (ref. 6) only around a week on the same cluster.

Structural alignment tools (reviewed in ref. 12) are slower for two reasons. First, whereas sequence search tools employ fast and sensitive prefilter algorithms to gain orders of magnitude in speed, no similar prefilters exist for structure alignment. Second, structural similarity scores are non-local: changing the alignment in one part affects the similarity in all other parts. Most structural aligners, such as the popular TM-align, Dali and CE[11,13,14], solve the alignment optimization problem by iterative or stochastic optimization.

To increase speed, a crucial idea is to describe the amino acid backbone of proteins as sequences over a structural alphabet and compare structures using sequence alignments[15]. Structural alphabets thus reduce structure comparisons to much faster sequence alignments. Many ways to discretize the local amino acid backbone have been proposed[16]. Most, such as CLE, 3D-BLAST and Protein Blocks, discretize the conformations of short stretches of usually 3–5 C_α atoms[17–19].

For Foldseek, we developed a type of structural alphabet that does not describe the backbone but, rather, tertiary interactions. The 20 states of the 3D interaction (3Di) alphabet describe for each residue *i* the geometric conformation with its spatially closest residue *j*. 3Di has three key advantages over traditional backbone structural alphabets. (1) Weaker dependency between consecutive letters and (2) more evenly distributed state frequencies, both enhancing information density and reducing false positives (FPs) (Supplementary Table 1). (3) The highest information density is encoded in conserved protein

[1]Quantitative and Computational Biology Group, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany. [2]School of Biological Sciences, Seoul National University, Seoul, South Korea. [3]Campus Institute Data Science (CIDAS), Göttingen, Germany. [4]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea. [5]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea. [6]These authors contributed equally: Michel van Kempen, Stephanie S. Kim. ✉e-mail: soeding@mpinat.mpg.de; martin.steinegger@snu.ac.kr

https://doi.org/10.1038/s41587-023-01773-0

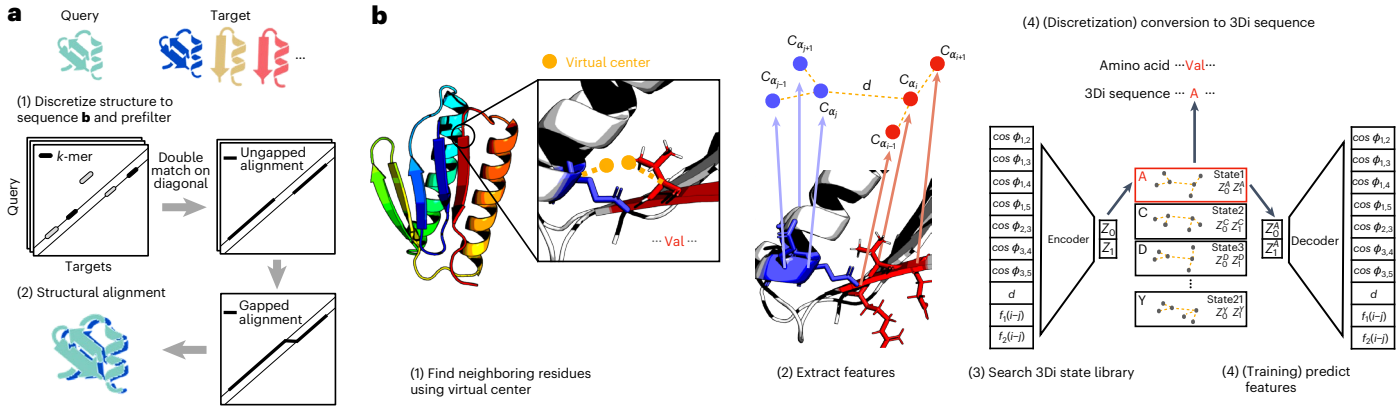


Fig. 1 | Foldseek workflow. a, Foldseek searches a set of query structures through a set of target structures. (1) Query and target structures are discretized into 3Di sequences (see **b**). To detect candidate structures, we apply the fast and sensitive *k*-mer and ungapped alignment prefilter of MMseqs2 to the 3Di sequences, (2) followed by vectorized Smith–Waterman local alignment combining 3Di and amino acid substitution scores. Alternatively, a global alignment is computed with a 1.7-times accelerated TM-align version (Supplementary Fig. 12). **b**, Learning the 3Di alphabet. (1) 3Di states describe tertiary interaction between a residue *i* and its nearest neighbor *j*. Nearest neighbors have the closest virtual center distance (yellow). Virtual center positions (Supplementary Fig. 1) were optimized for maximum search sensitivity. (2) To describe the interaction geometry of residues *i* and *j*, we extract seven angles, the Euclidean C_α distance and two sequence distance features from the six C_α coordinates of the two backbone fragments (blue and red). (3) These 10 features are used to define 20 3Di states by training a VQ-VAE[28] modified to learn states that are maximally evolutionary conserved. For structure searches, the encoder predicts the best-matching 3Di state for each residue.

cores and the lowest in non-conserved coil/loop regions, whereas the opposite is true for backbone structural alphabets.

Foldseek (https://foldseek.com/) (Fig. 1a) (1) discretizes the query structures into sequences over the 3Di alphabet and then uses a pre-trained 3Di substitution matrix (Supplementary Table 2) to search through the 3Di sequences of the target structures using the double-diagonal *k*-mer-based prefilter and gapless alignment prefilter modules from MMseqs2, our open-source sequence search software[6]. (2) High-scoring hits are aligned locally using 3Di (default) or globally with TM-align (Foldseek-TM). The local alignment stage combines 3Di and amino acid substitution scores. The construction of the 3Di alphabet is summarized in Fig. 1b and Supplementary Figs. 1–3.

To reduce high-scoring FPs and provide reliable *E* values, we subtracted the reversed query alignment score from the original score and applied a compositional bias correction within a local 40-residue sequence window (see the 'Pairwise local structural alignments' subsection). *E* values are calculated using an extreme-value score distribution, with parameters predicted by a neural network based on 3Di sequence composition and query length (see the '*E* values' subsection). Ranking of hits is determined by alignment bit score multiplied by the geometric mean of alignment TM-score and local distance difference test (LDDT). Foldseek also reports the probability for each match to be homologous, based on a fit of true and false matches on SCOPe.

We measured the sensitivity and speed of Foldseek, six protein structure alignment tools, an alignment-free structure search tool (Geometricus[20]) and a sequence search tool (MMseqs2 (ref. 6)) on the SCOPe dataset of manually classified single-domain structures[21]. Clustering SCOPe 2.01 at 40% sequence identity yielded 11,211 non-redundant protein sequences (SCOPe40). We performed an all-versus-all search and compared the tools' performance for finding members of the same SCOPe family, superfamily and fold (true-positive (TP) matches) by measuring for each query the fraction of TPs out of all possible correct matches until the first FP, a match to a different fold (see the 'SCOPe benchmark' subsection).

We first measured the sensitivity to detect relationships at family and superfamily level by the area under the curve (AUC) of the cumulative receiver operating characteristic (ROC) curve up to the first FP (Fig. 2a and Supplementary Fig. 4). Foldseek's sensitivity is below Dali and TM-align, higher than the structural aligner CE and much above the structural alphabet-based search tools 3D-BLAST and CLE-SW (Fig. 2a). In a precision-recall analysis, Foldseek-TM and Foldseek have the highest and third-highest area under the precision-recall curve on each of the three levels (Fig. 2b and Supplementary Fig. 4). Notably, Foldseek-TM improves over TM-align because its prefilter suppresses high-scoring FPs. Both sort hits by the average query and target length normalized TM-scores for best performance in the SCOPe benchmark.

Foldseek's performance is similar across all six secondary structure classes in SCOPe (Supplementary Fig. 5). On this small SCOPe40 benchmark set, Foldseek is more than 4,000 times faster than TM-align and Dali and over 21,000 times faster than CE (Fig. 2c). On the much larger AlphaFoldDB (version 1), where Foldseek approaches its full speed, it is around 184,600 and 23,000 times faster than Dali and TM-align, respectively (see below).

We devised a reference-free benchmark to assess search sensitivity and alignment quality of structural aligners (Fig. 2d) on a realistic set of full-length, multi-domain proteins. We clustered the AlphaFoldDB (version 1) to 34,270 structures using BLAST and SPICi[22]. We randomly selected 100 query structures from this set and aligned them against the remaining structures. TP matches are those with an LDDT score[23] of at least 0.6 and FPs below 0.25, ignoring matches in between. We set the LDDT thresholds according to the median inter-fold and intra-fold superfamily and family LDDT scores of SCOPe40 alignments (Supplementary Fig. 6). For other thresholds, see Supplementary Fig. 7. A domain-based sensitivity assessment would require a reference-based prediction of domains. To avoid it, we evaluated the sensitivity per residue. Figure 2d shows the distribution of the fraction of query residues that were part of alignments with at least *x* TP targets with better scores than the first FP match. Again, Foldseek has similar sensitivity as Dali, CE and TM-align and much higher sensitivity than CLE-SW and MMseqs2.

We analyzed the quality of alignments produced by the top five matches per query. We computed the alignment sensitivity as the number of TP residues divided by the query length and the precision as the number of TP residues divided by the alignment length. TP residues are those with residue-specific LDDT score above 0.6; FP residues are below 0.25; and residues with other scores are ignored. Figure 2e shows the average sensitivity versus precision of the 100 × 5 structure alignments.

https://doi.org/10.1038/s41587-023-01773-0

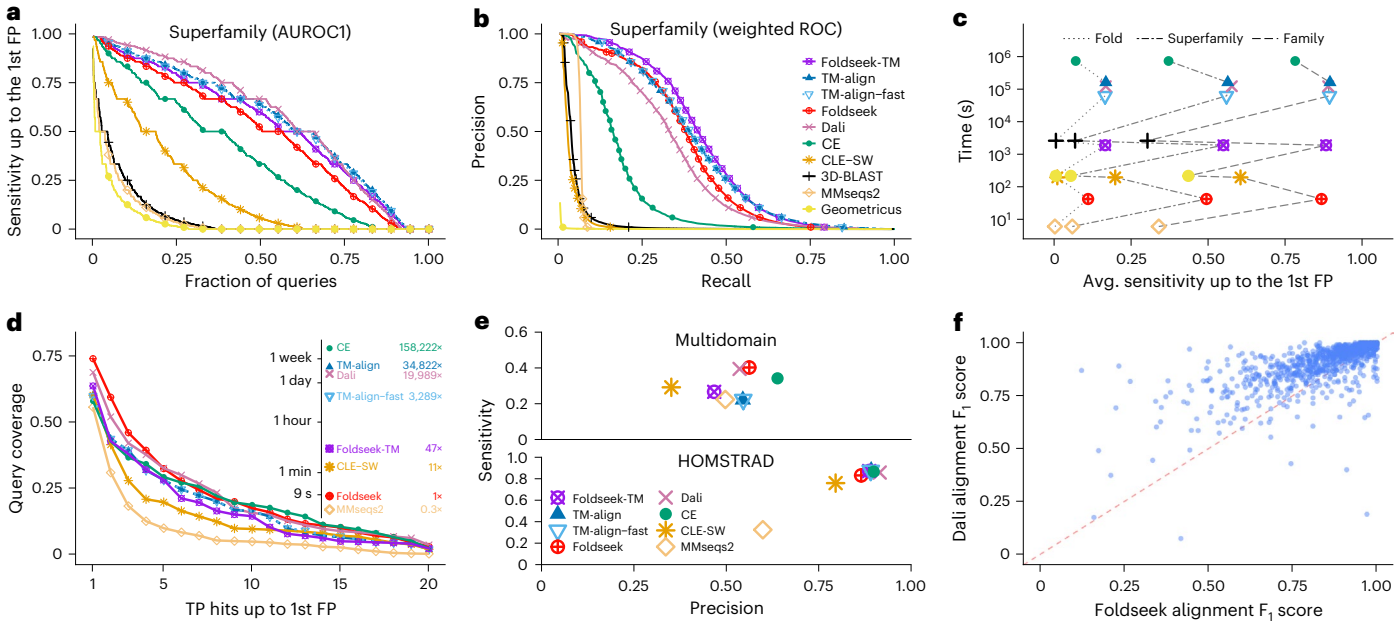


Fig. 2 | Foldseek reaches similar sensitivities as structural aligners at thousands of times their speed. a, Cumulative distributions of sensitivity for homology detection on the SCOPe40 database of single-domain structures. TPs are matches within the same superfamily; FPs are matches between different folds. Sensitivity is the area under the ROC (AUROC) curve up to the first FP (see Supplementary Fig. 4 for family and fold). **b**, Precision-recall curve of SCOPe40 superfamilies (see Supplementary Fig. 4 for family and fold). **c**, Average sensitivity up to the first FP for family, superfamily and fold versus total runtime on an AMD EPYC 7702P 64-core CPU for the all-versus-all searches of 11,211 structures of SCOPe40. **d**, Search sensitivity on multi-domain, full-length AlphaFold2 protein models. One hundred queries, randomly selected from AlphaFoldDB (version 1), were searched against this database. Per-residue query coverage (*y* axis) is the fraction of residues covered by at least *x* (*x* axis) TP matches ranked before the first FP match. **e**, Alignment quality for alignments of AlphaFoldDB (version 1) protein models (top panel), averaged over the top five matches of each of the 100 queries. Sensitivity = TP residues in alignment / query length; precision = TP residues / alignment length. Reference-based alignment quality benchmark on HOMSTRAD alignments. **f**, Alignment quality comparison between Foldseek and Dali for each HOMSTRAD family. The F_1 score is the harmonic mean between sensitivity and precision.

Foldseek alignments are more accurate and sensitive than MMseqs2, CLE-SW and TM-align, similarly accurate as Dali and 13% less precise but 15% more sensitive than CE. In the reference-based HOMSTRAD alignment quality benchmark[24], Foldseek performs slightly below CE, Dali and TM-align (Fig. 2e). Figure 2f shows the comparison between Foldseek and Dali in alignment quality for all HOMSTRAD families (see Supplementary Fig. 8 for example alignments).

To find potentially problematic high-scoring Foldseek FPs, we searched the set of unfragmented models in AlphaFoldDB (version 1) with average predicted LDDT[1]≥80 against itself. We inspected the 1,675 (of 133,813) high-scoring FPs (score per aligned column ≥ 1.0, TM-score < 0.5), revealing queries with multiple structured segments but with incorrect relative orientations (Supplementary Table 3 and Supplementary Fig. 9). The folded segments were correctly aligned by Foldseek. This illustrates that 3D aligners such as TM-align may overlook homologous structures that are not globally superposable, whereas Foldseek (as well as the two-dimensional (2D) aligner Dali) is independent of relative domain orientations and excels at detecting homologous multi-domain structures[12].

We developed a webserver (https://search.foldseek.com) for multi-database searches, including AlphaFoldDB (version 4: Proteomes and Swiss-Prot), AlphaFoldDB (version 4) and CATH[25] clustered at 50% sequence identity, ESM Atlas-HQ and Protein Data Bank (PDB)[26].

We compared Foldseek webserver, TM-align and Dali using SARS-CoV-2 RdRp (PDB: 6M71, chain A (ref. 27); 942 residues) in AlphaFoldDB (version 1). Search times were 10 d for Dali, 33 h for TM-align and 6 s for Foldseek, making it 180,000 and 23,000 times faster. All top 10 hits were known RdRp homologs (Supplementary Table 4).

The availability of high-quality structures for nearly every folded protein is transformative for biology and bioinformatics. Sequence-based analyses will soon be largely superseded by structure-based analyses. The main limitation in our view—the four orders of magnitude slower speed of structure comparisons—is removed by Foldseek.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01773-0.

References

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
3. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein–sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
4. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
5. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
6. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

7. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
8. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
9. Mahlich, Y., Steinegger, M., Rost, B. & Bromberg, Y. HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics* **34**, i304–i312 (2018).
10. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).
11. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
12. Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.* **19**, 341–348 (2009).
13. Holm, L. Using Dali for protein structure comparison. *Methods Mol. Biol.* **2112**, 29–42 (2020).
14. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747 (1998).
15. Guyon, F., Camproux, A.-C., Hochez, J. & Tuffery, P. SA-Search: a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res.* **32**, W545–W548 (2004).
16. Ma, J. & Wang, S. Algorithms, applications, and challenges of protein structure alignment. *Adv. Protein Chem. Struct. Biol.* **94**, 121–175 (2014).
17. Wang, S. & Zheng, W.-M. CLePAPS: fast pair alignment of protein structures based on conformational letters. *J. Bioinform. Comput. Biol.* **6**, 347–366 (2008).
18. Yang, J.-M. & Tung, C.-H. Protein structure database search and evolutionary classification. *Nucleic Acids Res.* **34**, 3646–3659 (2006).
19. de Brevern, A. G., Etchebest, C. & Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41**, 271–287 (2000).
20. Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics* **36**, i718–i725 (2020).
21. Chandonia, J.-M., Fox, N. K. & Brenner, S. E. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res.* **47**, D475–D481 (2019).
22. Jiang, P. & Singh, M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics* **26**, 1105–1111 (2010).
23. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
24. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471 (1998).
25. Bordin, N. et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun. Biol.* **6**, 160 (2023).
26. Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
27. Gao, Y. et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).
28. Van den Oord, A., Vinyals, O. & Kavukcuoglu, K. Neural discrete representation learning. *Proc. of the 31st Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf (NIPS, 2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Overview

Foldseek enables fast and sensitive comparison of large structure sets. It encodes structures as sequences over the 20-state 3Di alphabet and, thereby, reduces structural alignments to 3Di sequence alignments. The 3Di alphabet developed for Foldseek describes tertiary residue–residue interactions instead of backbone conformations and proved critical for reaching high sensitivities. Foldseek's prefilter finds two *similar*, spaced 3Di *k*-mer matches in the same diagonal of the dynamic programming matrix. By not restricting itself to exact matches, the prefilter achieves high sensitivity while reducing the number of sequences for which full alignments are computed by several orders of magnitude. Further speed-ups are achieved by multi-threading and using single instruction, multiple data (SIMD) vector units. Owing to the SIMDe library (https://github.com/simd-everywhere/simde), Foldseek runs on a wide range of CPU architectures (x86_64, arm64 and ppc64le) and operating systems (Linux and macOS). The core modules of Foldseek, which build on the MMseqs2 framework[6], are described in the following paragraphs.

Create database

The `createdb` module converts a set of PDB (ref. 29), macromolecular crystallographic information file (mmCIF) formatted files or Foldcomp compressed structure (FCZ (ref. 30)) files into an internal Foldseek database format using the Gemmi package (https://gemmi.readthedocs.io/en/latest/) or the Foldcomp library. The format is compatible with the MMseqs2 database format, which is optimized for parallel access. We store each chain as a separate entry in the database. The module follows the MMseqs2 `createdb` module logic. However, in addition to the amino acid sequence, it computes the 3Di sequence from the 3D atom coordinates of the backbone atom and C_β coordinates (see the 'Descriptors for 3Di structural alphabet' and 'Optimize nearest-neighbor selection' subsections). Backbone atom and C_β coordinates are needed only for the nearest-neighbor selection. For C_α-only structures, Foldseek reconstructs backbone atom coordinates using PULCHRA[31]. Missing C_β coordinates (for example, in glycines) are defined such that the four groups attached to the C_α are arranged at the vertices of a regular tetrahedron. The 3Di and amino acid sequences and the C_α coordinates are stored in the Foldseek database. To save disk space, we optionally compress the C_α coordinates losslessly, beginning with three uncompressed 4-byte floating-point C_α coordinates and storing all subsequent coordinates as 2-byte signed integer differences[32]. If any difference is too large to be represented with a 2-byte signed integer, we fall back to 4-byte floats for all C_α coordinates.

Prefilter

The `prefilter` module detects double matches of similar, spaced words (*k*-mers) that occur on the same diagonal. The *k*-mer size is dynamically set to $k = 6$ or $k = 7$ depending on the size of the target database. Similar *k*-mers are those with a 3Di substitution matrix score above a certain threshold, whereas MMseqs2 uses an amino acid substitution matrix to compute the similarity (see the '3Di substitution score matrix' subsection). The gapless double-match criterion suppresses hits to non-homologous structures effectively, as they are less likely to have consecutive *k*-mer matches on the same diagonal by chance. To avoid FP matches due to regions with biased 3Di sequence composition, a compositional bias correction is applied in a way analogous to MMseqs2 (ref. 33). For each hit, we perform an ungapped alignment over the diagonals with double, consecutive, similar *k*-mer matches and sort those by the maximum ungapped diagonal score. Alignments with a score of at least 15 bits are passed on to the next stage. We implemented an optional taxonomy filter within the prefiltering step to help users search through taxonomic subsets of the target database. After the gapless double-diagonal matching stage and before the ungapped alignment stage, we reject all potential target hits that do not lie within a taxonomic clade specified by the user.

Pairwise local structural alignments

After the prefilter has removed the vast majority of non-homologous sequences, the `structurealign` module computes pairwise alignments for the remaining sequences using an SIMD-accelerated Smith–Waterman algorithm[34,35]. We extended this implementation to support amino acid and 3Di scoring, compositional bias correction and 256-bit-wide vectorization. The score linearly combines amino acid and 3Di substitution scores with weights 1.4 and 2.1, respectively. We optimized these two weights and the ratio of gap extend to gap open penalty on ~1% of alignments (all-versus-all on 10% of randomly selected SCOPe40 domains). A compositional bias correction is applied to the amino acid and 3Di scores. To further suppress high-scoring FP matches, for each match we align the reversed query sequence against the target and subtract the reverse bit score from the forward bit score.

Structural bit score

We rank hits by a 'structural bit' score—that is, the product of the bit score produce by the Smith–Waterman algorithm and the geometric mean of average alignment LDDT and the alignment TM-score.

Fast alignment LDDT computation

To improve the LDDT score computation speed, we store the 3D coordinates of the query in a grid using spatial hashing. Each grid cell spans 15 Å, which is the default radius considered for the LDDT computation. For each aligned query residue *i*, we compute the distances to all C_α atoms within a 15 Å radius by searching all neighboring grid cells of the query residue's grid cell. For each residue *j*, we compute the distance between the C_α atoms of *i* and *j* and the distance of the corresponding aligned target residues. Query and target distances for the aligned pairs are subtracted, and the differences *d* are transformed into LDDT scores $s = 0.25 \times ((d < 0.5) + (d < 1.0) + (d < 2.0) + (d < 4.0))$. For each *i*, we obtain the means of the scores for all Cα atoms *j* within the 15 Å radius of *i*. The LDDT score is the mean of these means over all query residues *i*.

E values

To estimate *E* values for each match, we trained a neural network to predict the mean μ and scale parameter λ of the extreme value distribution for each query. The module `computemulambda` takes a query and database structures as input and aligns the query against a randomly shuffled version of the database sequences. For each query sequence, the module produces *N* random alignments and fits to their scores an extreme value (Gumbel) distribution. The maximum likelihood fitting is done using the Gumbel fitting function taken from HMMER3 (`hmmcalibrate`)[36]. To train the neural network, it is critical to use query and target proteins that include problematic regions, such as structurally biased, disordered or badly modeled regions that occur ubiquitously in full-length proteins or modeled structures. We, therefore, trained the network on 100,000 structures sampled from the AlphaFoldDB (version 1). We trained a neural network to predict μ and λ from the amino acid composition of the query and its length (so a scrambled version of the query sequence would produce the same μ and λ). The network has 22 input nodes, two fully connected layers with 32 nodes each (ReLU activation) and two linear output nodes. The optimizer Adam with learning rate 0.001 was used for training. When testing the resulting *E* values on searches with scrambled sequences, the log of the mean number of FPs per query turned out to have an accurately linear dependence on the log of the reported *E* values, albeit with a slope of 0.32 instead of 1. We, therefore, correct the *E* values from the neural network by taking them to the power of 0.32. We compared how well the mean number of FPs at a given *E* value agreed with the *E* values reported by Foldseek, MMseqs2 and 3D-Blast (Supplementary Fig. 10; see Supplementary Fig. 11 for AlphaFoldDB). We considered a hit as FP if it was in a different fold and had a TM-score lower than 0.3. Furthermore, we ignored all cross-fold hits within the four-bladed to eight-bladed β-propeller superfamilies (SCOPe b.66-b.70) and within

the Rossman-like folds (c.2–c.5, c.27, c.28, c.30 and c.31) because of the extensive cross-fold homologies within these groups[37].

Probability of TP match

Foldseek computes for each match a simple estimate for the probability that the match is a TP match given its structural bit score. Here, hits within the same superfamily are TP; hits to another fold are FP; and hits to the same family or to another superfamily are ignored. We estimate the structural bit score distributions of TP and FP hits ($p(score|TP)$ and $p(score|FP)$), which allow us to calculate the probability of a TP $p(TP|score) = \frac{p(\text{score}|TP)p(TP)}{p(\text{score}|TP)p(TP)+p(\text{score}|FP)p(FP)}$. Both score distributions were fitted on SCOPe40 with a mixture model consisting of two gamma distributions (resulting in five parameters for each function). For the fitting, the function `gammamixEM` from the R package `mixtools`[38] was used. We excluded cross-fold hits between certain folds as in the *E* value estimation. For example, Foldseek finds around the same number of FPs and TPs with a score of 51 in SCOPe40. The probability for a hit with score 51 is, therefore, 50%.

Pairwise global structural alignments using TM-align

We also offer the option to use TM-align for pairwise structure alignment instead of the 3Di-based alignment. We implemented TM-align based on the C_α atom coordinates and made adjustments to improve the (1) speed and (2) memory usage. (1) TM-align performs multiple floating-point-based Needleman–Wunsch (NW) alignment steps while applying different scoring functions (for example, score secondary structure, Euclidean distance of superposed structures or fragments). TM-align's NW code did not take advantage of SIMD instructions; therefore, we replaced it by parasail's[39] SIMD-based NW implementation and extended it to support the different scoring functions. We also replaced the TM-score computation using fast_protein_cluster's SIMD-based implementation[40]. Our NW implementation does not compute exactly the same alignment because we apply affine gap costs, whereas TM-align does not (Supplementary Fig. 12). (2) TM-align requires 17 bytes × query length × target length of memory, and we reduce the constant overhead from 17 bytes to 4 bytes. If Foldseek is used in TM-align mode (parameter `--alignment-type 1`), TM-align is used for the alignment stage after the prefilter step, where we replace the reported *E* value column with TM-scores normalized by the query length. The results are ordered in descending order by average TM-score by default.

Descriptors for 3Di structural alphabet

The 3Di alphabet describes the tertiary contacts between residues and their nearest neighbors in 3D space. For each residue *i*, the conformation of the local backbone around *i*, together with the local backbone around its nearest neighbor *j*, is approximated by 20 discrete states (Supplementary Fig. 3). We chose the alphabet size $A = 20$ as a tradeoff between encoding as much information as possible (large *A*; Supplementary Fig. 13) and limiting the number of similar 3Di *k*-mers that we need to generate in the *k*-mer-based prefilter, which scales with A^k. The discrete single-letter states are formed from neighborhood descriptors containing 10 features encoding the conformation of backbones around residues *i* and *j* represented by the C_α atoms ($C_{\alpha,i-1}$, $C_{\alpha,i}$, $C_{\alpha,i+1}$) and ($C_{\alpha,j-1}$, $C_{\alpha,j}$, $C_{\alpha,j+1}$). The descriptors use the five unit vectors along the following directions:

$$\mathbf{u_1}: C_{\alpha,i-1} \to C_{\alpha,i} \quad \mathbf{u_4}: C_{\alpha,j} \to C_{\alpha,j+1}$$
$$\mathbf{u_2}: C_{\alpha,i} \to C_{\alpha,i+1} \quad \mathbf{u_5}: C_{\alpha,i} \to C_{\alpha,j}$$
$$\mathbf{u_3}: C_{\alpha,j-1} \to C_{\alpha,j}.$$

We define the angle between u_k and u_l as ϕ_{kl}, so $\cos\phi_{kl} = \mathbf{u_k^T u_l}$. The seven features $\cos\phi_{12}, \cos\phi_{34}, \cos\phi_{15}, \cos\phi_{35}, \cos\phi_{14}, \cos\phi_{23}, \cos\phi_{13}$ and the distance $|C_{\alpha,i} - C_{\alpha,j}|$ describe the conformation between the backbone fragments. In addition, we encode the sequence distance with the two features $\text{sign}(i-j)\ \min(|i-j|, 4)$ and $\text{sign}(i-j)\ \log(|i-j|+1)$.

Learning the 3Di states using a vector quantized variational autoencoder

The 10-dimensional descriptors were discretized into an alphabet of 20 states using a vector quantized variational autoencoder (VQ-VAE)[28]. In contrast to standard clustering approaches such as *k*-means, VQ-VAE is a nonlinear approach that can optimize decision surfaces for each of its states. In contrast to the standard VQ-VAE, we trained the VQ-VAE not as a simple generative model but, rather, to learn states that are maximally conserved in evolution. To that end, we trained it with pairs of descriptors $\mathbf{x}_n, \mathbf{y}_n \in \mathbb{R}^{10}$ from structurally aligned residues, to predict the distribution of y_n from x_n.

The VQ-VAE consists of an encoder and decoder network with the discrete latent 3Di state as a bottleneck in between. The encoder network embeds the 10-dimensional descriptor x_n into a 2D continuous latent space, where the embedding is then discretized by the nearest centroid, each centroid representing a 3Di state. Given the centroid, the decoder predicts the probability distribution of the descriptor y_n of the aligned residue. After training, only encoder and centroids are used to discretize descriptors. Encoder and decoder networks are both fully connected with two hidden layers of dimension 10, a batch normalization after each hidden layer and ReLU as activation functions. The encoder, centroids and decoder have 242, 40 and 352 parameters, respectively. The output layer of the decoder consists of 20 units predicting μ and σ^2 of the descriptors *x* of the aligned residue, such that the decoder predicts $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{l\sigma^2})$ (with diagonal covariance).

We trained the VQ-VAE on the loss function defined in Equation (3) in ref. 28 (with commitment loss = 0.25) using the deep learning framework PyTorch (version 1.9.0), the Adam optimizer, with a batch size of 512, and a learning rate of 10^{-3} over four epochs. Using Kerasify (https://github.com/moof2k/kerasify), we integrated the encoder network into Foldseek. The domains from SCOPe40 were split 80%/20% by fold into training and validation sets. For the training, we aligned the structures with TM-align, removed all alignments with a TM-score below 0.6 and removed all aligned residue pairs with a distance between their C_α atoms of more than 5 Å. We trained the VQ-VAE with 100 different initial parameters and chose the model that was performing best in the benchmark on the validation dataset (the highest sum of ratios between 3Di AUC and TM-align AUC for family, superfamily and fold level).

3Di substitution score matrix

We trained a BLOSUM-like substitution matrix for 3Di sequences from pairs of structurally aligned residues used for the 'VQ-VAE training'. First, we determined the 3Di states of all residues. Next, the substitution frequencies among 3Di states were calculated by counting how often two 3Di states were structurally aligned. (Note that the substitution frequencies from state A to state B and the opposite direction are equal.) Finally, the score $S(x,y) = 2\log_2\frac{p(x,y)}{p(x)p(y)}$ for substituting state x through state y is the log-ratio between the substitution frequency $p(x, y)$ and the probability that the two states occur independently, scaled by the factor 2.

3Di alphabet cross-validation

We trained the 3Di alphabet (the VQ-VAE weights) and the substitution matrix by four-fold cross-validation on SCOPe40. We split the SCOPe40 dataset into four parts, such that all domains of each fold ended up in the same part of the four parts. 3Di alphabets were trained on three parts and tested on the remaining part, selecting each of the four parts in turn as a test set. The 80:20 split between training and validation sets to select the best alphabet out of the 100 VQ-VAE runs happens within the 3/4 of the cross-validation training data. Supplementary Fig. 14 shows the mean sensitivity (black) and the standard deviation (gray

area) in comparison to the final 3Di alphabet, for which we trained the 3Di alphabet on the entire SCOPe40 (red). No overfitting was observed, despite training 492 parameters (282 neural network and 210 substitution matrix entries). In Fig. 2, we, therefore, show the benchmark results for the final 3Di alphabet, trained on the entire SCOPe40.

Nearest-neighbor selection

To select nearest-neighbor residues that maximize the performance of the resulting 3Di alphabet in finding and aligning homologous structures, we introduced the virtual center V of a residue. The virtual center position is defined by the angle θ (V-C_α-C_β), the dihedral angle τ (V-C_α-C_β-N) and the length l ($|V - C_\alpha|$) (Supplementary Fig. 1). For each residue i, we selected the residue j with the smallest distance between their virtual centers. The virtual center was optimized on the training and validation structure sets used for the VQ-VAE training by creating alphabets for positions with $\theta \in [0, 2\pi]$, $\tau \in [-\pi, \pi]$ in 45° steps and $l \in \{1.53\text{ Å } k\text{: } k \in \{1, 1.5, 2, 2.5, 3\}\}$ (1.53 Å is the distance between C_α and C_β). The virtual center defined by $\theta = 270°$, $\tau = 0°$ and $l = 2$ performed best in the SCOPe benchmark.

This virtual center preferably selects long-range, tertiary interactions and only falls back to selecting interactions to $i + 1$ or $i - 1$ when no other residues are nearby. In that case, the interaction captures only the backbone conformation.

SCOPe benchmark

We downloaded the SCOPe40 structures (available at https://wwwuser.gwdg.de/~compbiol/foldseek/scop40pdb.tar.gz).

The SCOPe benchmark set consists of single domains with an average length of 174 residues. In our benchmark, we compare the domains all-versus-all. Per domain, we measured the fraction of detected TPs up to the first FP. For family-level, superfamily-level and fold-level recognition, TPs were defined as same family, same superfamily and not same family and same fold and not same superfamily, respectively. Hits from different folds are FPs.

Evaluation SCOPe benchmark

After sorting the alignment result of each query (described in the 'Tools and options for benchmark comparison' subsection), we calculated the sensitivity as the fraction of TPs in the sorted list up to the first FP, all excluding self-hits. For comparison, we took the mean sensitivity over all queries for family-level, superfamily-level and fold-level classifications. We evaluated only SCOPe members with at least one other family, superfamily and fold member. We measure the sensitivity up to the 1st FP (ROC1) instead, for example, up to the 5th FP, because ROC1 better reflects the requirements for low false discovery rates in automatic searches.

Additionally, we plotted precision-recall curves for each tool (Fig. 2b and Supplementary Fig. 4). After sorting the alignment results by the structural similarity scores (as described in the 'Tools and options for benchmark comparison' subsection) and excluding self-matches, we generated a weighted precision-recall curve for family-level, superfamily-level and fold-level classifications (precision = TP / (TP + FP) and recall = TP / (TP + FN)). All counts (TP, FP and FN) were weighted by the reciprocal of their family, superfamily or fold size. In this way, folds, superfamilies and families contribute linearly with their size instead of quadratically[36].

Runtime evaluations on SCOPe and AlphaFoldDB

We measured the speed of structural aligners on a server with an AMD EPYC 7702P 64-core CPU and 1,024 GB RAM memory. On SCOPe40, we measured or estimated the runtime for an all-versus-all comparison. To avoid excessive runtimes for TM-align, Dali and CE, we estimated the runtime by randomly selecting 10% of the 11,211 SCOPe domains as queries. We measured runtimes on AlphaFoldDB for searches with the same 100 randomly selected queries used for the sensitivity and alignment quality benchmarks (Fig. 2d,e). Tools with multi-threading support (MMseqs2 and Foldseek) were executed with 64 threads; tools without were parallelized by breaking the query set into 64 equally sized chunks and executing them in parallel.

Reference-free multi-domain benchmarks

We devised two reference-free benchmarks that do not rely on any reference structural alignments. We clustered the AlphaFoldDB (version 1)[3] using SPICi[22]. For this, we first aligned all protein sequences all against all using an E value threshold $<10^{-3}$ using BLAST (2.5.0+)[5]. SPICi produced 34,270 clusters from the search result. For each cluster, we picked the longest protein as representative. We randomly selected 100 representatives as queries and searched the set of remaining structures. The top five alignments of all queries are listed at https://wwwuser.gwdg.de/~compbiol/foldseek/multi_domain_top5_alignments/.

For the evaluation, we needed to adjust the LDDT score function taken from AlphaFold2 (ref. 1). LDDT calculates for each residue i in the query the fraction of residues in the 15 Å neighborhood that have a distance within 0.5, 1, 2 or 4 Å of the distance between the corresponding residues in the target[23]. The denominator of the fraction is the number of 15 Å neighbors of i that are aligned to some residue in the target. This does not properly penalize non-compact models in which each residue has few neighbors within 15 Å. We, therefore, use as denominator the *total* number of neighboring residues within 15 Å of i.

For the alignment quality benchmark (Fig. 2e), we classified each aligned residue pair as TP or FP depending on its residue-wise LDDT score—that is, the fraction of distances to its 15 Å neighbors that are within 0.5, 1, 2 and 4 Å of the distance to the corresponding residues in the query, averaged over the four distance thresholds. TP residues are those with a residue-wise LDDT score of at least 0.6 and FPs below 0.25, ignoring matches in between. For the search sensitivity benchmark (Fig. 2d), TP residue–residue matches are those with an LDDT score of the query-target alignment of at least 0.6 and FPs below 0.25, ignoring matches in between. (The LDDT score of the query-target alignment is the average of the residue-wise LDDT score over all aligned residue pairs.) The choice of thresholds is illustrated in Supplementary Fig. 6. The benchmark for other thresholds is shown in Supplementary Fig. 7.

All-versus-all search of AlphaFoldDB with Foldseek

We downloaded the AlphaFoldDB (version 1)[3] containing 365,198 protein models and searched it all-versus-all using Foldseek `-s 9.5 --max-seqs 2000`. For our second-best hit analysis, we consider only models with (1) an average C_α's predicted LDDT (pLDDT) greater than or equal to 80 and (2) models of non-fragmented domains. We also computed the structural similarity for each pair using TM-align (default options).

Tools and options for benchmark comparison

Owing to dataset overlap, we excluded methods from the benchmark that are likely to be overfitted on SCOPe. This applies to methods that trained many thousands of parameters (for example, deep neural networks) with strong data leakage among training, validation and test sets. For example, several tools allowed up to 40% sequence identity between sets. The following command lines were used in the SCOPe as well as the multi-domain benchmark:

Foldseek

We used Foldseek commit `aeb5e` during this analysis. Foldseek was run with the following parameters: `--threads 64 -s 9.5 -e 10 --max-seqs 2000`.

Foldseek-TM

For the Foldseek-TM benchmark, we first run a regular 3Di/AA-based Foldseek search using the following parameters: `--threads 64 -s 9.5 -e 10 --max-seqs 4000 --alignment-mode 1`. All hits

passing are then aligned by Foldseeks's `tmalign --tmalign-fast 1 --tmscore-threshold 0.0 -a`. We used Foldseek commit `aeb5e` during this analysis. We expose Foldseek-TM in our command-line interface as a search mode that combines regular Foldseek 3Di/AA-based workflow with our TM-align implementation within the `tmalign` module.

MMseqs2

We used the default MMseqs2 (release `13-45111`) search algorithm to obtain the sequence-based alignment result. MMseqs2 sorts the results by *E* value and score. We searched with: `--threads 64 -s 7.5 -e 10000 --max-seqs 2000.`

CLE-Smith–Waterman

We used PDB Tool version 4.80 (https://github.com/realbigws/PDB_Tool) to convert the benchmark structure set to CLE sequences. After the conversion, we used SSW[35] (commit `ad452e`) to align CLE sequences all-versus-all. We sorted the results by alignment score. The following parameters were used to run SSW: (1) protein alignment mode (`-p`); (2) gap open penalty of 100 (`-o 100`); (3) gap extend penalty of 10 (`-e 10`); (4) CLE's optimized substitution matrix (`-a cle.shen.mat`); and (5) returning alignment (`-c`). The gap open and extend values were inferred from DeepAlign[41]. The results are sorted by score in descending order.

```
ssw_test -p -o 100 -e 10 -a cle.shen.mat -c
```

3D-BLAST

We used 3D-BLAST (beta102) with BLAST+ (2.2.26) and SSW[34] (version ad452e). We first converted the PDB structures to a 3D-BLAST database using `3d-blast -sq_write` and `3d-blast -sq_append`. We searched the structural sequences against the database using `blastp` with the following parameters: (1) 3D-BLAST's optimized substitution matrix (`-M 3DBLAST`); (2) number of hits and alignments shown of 12,000 (`-v 12000 -b 12000`); (3) *E* value threshold of 1,000 (`-e 1000`); (4) disabling query sequence filter (`-F F`); (5) gap open of 8 (`-G 8`); and (6) gap extend of 2 (`-E 2`). 3D-BLAST's results are sorted by *E* value in ascending order:

```
blastall -p blastp -M 3DBLAST -v 12000 -b 12000 -e
1000 -F F -G 8 -E 2
```

For Smith–Waterman, we used (1) gap open of 8; (2) gap extend of 2; (3) returning alignments (`-c`); (4) 3D-BLAST's optimized substitution matrix (`-a 3DBLAST`); and (5) protein alignment mode (`-p`): `ssw_test -o 8 -e 2 -c -a 3DBLAST -p`. We noticed that the 3D-BLAST matrix with Smith–Waterman resulted in a similar performance to CLE: 0.717, 0.230 and 0.011 for family classification, superfamily classification and fold classification, respectively. We excluded 3D-BLAST's measurement from the multi-domain benchmark because it produced occasionally high scores (>10^7) for single residue alignments.

TM-align

We downloaded and compiled the `TMalign.cpp` source code (version 2019/08/22) from the Zhang group website. We ran the benchmark using default parameters and `-fast` for the fast version. TM-align reports two TM-scores: (1) normalized by the length of 1st chain (query) or (2) normalized by the length of the 2nd chain (target). We used the average of TM-scores normalized by the 1st chain (query) and 2nd chain (target) in all our analyses. We evaluated TM-align's performance by sorting the results based on both the query TM-score and the minimum, maximum and average TM-score for both the query and target. Our results showed that the average TM-score performed the best in our single-domain benchmark.

Default: `TMalign query.pdb target.pdb`
Fast: `TMalign query.pdb target.pdb -fast`

Dali

We installed the standalone DaliLite.v5. For the SCOPe40 benchmark set, input files were formatted in DAT files with Dali's `import.pl`. The conversion to DAT format produced 11,137 valid structures out of the 11,211 initial structures for the SCOPe benchmark and 34,252 structures out of 34,270 SPICi clusters. After formatting the input files, we calculated the protein alignment with Dali's structural alignment algorithm. The results were sorted by Dali's *z*-score in descending order:

```
import.pl -pdbfile query.pdb -pdbid PDBid -dat DAT
dali.pl -cd1 queryDATid -db targetDB.list -TITLE
systematic -dat1 DAT -dat2 DAT -outfmt "summary"
-clean
```

CE

We used BioJava's[42] (version 5.4.0) implementation of the combinatorial extension (CE) alignment algorithm. We modified one of the modules of BioJava under shape configuration to calculate the CE value. Our modified `CEalign.jar` file requires a list of query files, path to the target PDB files and an output path as input parameters. This Java module runs an all-versus-all CE calculation with unlimited gap size (`maxGapSize -1`) to improve alignment results[14]. The results were sorted by *z*-score in descending order. For the multi-domain benchmark, we excluded one query that was running over 16 d. The Jar file of our implementation of CE calculation is provided (see 'Code availability').

```
java -jar CEalign.jar querylist.txt
TargetPDBDirectory OutputDirectory
```

Geometricus

We included Geometricus[20] in the SCOPe benchmark as a representative of alignment-free tools, which are fast but can find only globally similar structures. Geometricus discretizes fixed-length backbone fragments (shape-mers) using their 3D moment invariants and represents structures as a fixed-length count vector over the shape-mers. To calculate the shape-mer-based structural similarity of the benchmark set, we used Caretta-shape's Python implementation (1e3adb0) of multiple structure alignment (https://github.com/TurtleTools/caretta/caretta/multiple_alignment.py), which computes the Bray–Curtis similarity between the Geometricus shape-mer vectors. Our modified version extracts structural information from the input files and generates all-versus-all pairwise structural similarity score as an output. We ran Geometricus on a single core because it would require substantial engineering efforts to implement parallelization on multiple cores. With an efficient multi-core implementation, Geometricus might be as fast as MMseqs2 on 64 cores. The Python code of our implementation of Geometricus is provided:

```
python runGeometricus_caretta.py -i querylist.txt
-o OutputDirectory
```

HOMSTRAD alignment benchmark

The HOMSTRAD database contains expert-curated homologous structural alignments for 1,032 protein families[24]. We downloaded the latest HOMSTRAD version (https://mizuguchilab.org/homstrad/data/homstrad_with_PDB_2022_Aug_1.tar.gz) and picked the pairwise alignments between the first and last members of each family, which resulted in structures of a median length of 182 residues. We used the same parameters as in the SCOPe and multi-domain benchmark. We forced Foldseek, MMseqs2 and CLE-Smith–Waterman to return an alignment by switching off the prefilter and *E* value threshold. With the HOMSTRAD alignments as reference, we measured for each pairwise alignment the sensitivity (fraction of residue pairs of the HOMSTRAD alignment that were correctly aligned) and the precision (fraction of correctly aligned residue pairs in the predicted alignment). Dali, CE

and CLE-Smith–Waterman failed to produce an alignment for 35, 1 and 1 out of 1,032 pairs, respectively, which were rated with a sensitivity of 0. The mean sensitivity and precision are shown in Fig. 2e, and all individual alignments are listed in `homstrad_alignments.txt` at https://wwwuser.gwdg.de/~compbiol/foldseek/.

Limitations of benchmarks

The SCOPe benchmark to measure search sensitivity uses only single-domain proteins as queries and targets (Fig. 2a–c). It, therefore, cannot assess the ability of tools to find local similarities—for example, finding homologous domains shared between two multi-domain proteins. The alignment benchmark based on HOMSTRAD (Fig. 2e) has the same limitation, as the vast majority of proteins in HOMSTRAD have a single domain (median length = 182 residues). A drawback of our reference-free multi-domain benchmark is the need to choose thresholds for TPs and FPs (Supplementary Fig. 6).

Pre-built and ready-to-download databases

Foldseek includes the `databases` module to aid users with the download and setup of structural databases. Currently, we include the four variants of the AlphaFoldDB (version 4): UniProt (214 million structures), UniProt50, a clustered database to 50% sequence identity and 90% bi-directional coverage using MMseqs2 (parameters `-c 0.9 --min-seq-id 0.5 --cluster-reassign`; 54 million structures), Proteome (564,000 structures) and Swiss-Prot (542,000 structures). Additionally, we regularly build and offer a 100% sequence identity clustered PDB. The update pipeline is available in the `util/update_webserver_pdb` folder in the main Foldseek repository. These databases are hosted on Cloudflare R2 for fast downloading. We additionally link to and provide an automatic setup procedure for the ESM Atlas High-Quality Clu30[4] database.

Webserver

The Foldseek webserver is based on the MMseqs2 webserver[43]. To allow for searches in seconds, we implemented MMseqs2's pre-computed database indexing capabilities in Foldseek. Using these, the search databases can be fully cached in system memory by the operating system and instantly accessed by each Foldseek process, thus avoiding expensive accesses to slow disk drives. A similar mechanism was used to store and read the associated taxonomic information. The AlpaFoldDB/UniProt50 (version 4), AlphaFoldDB/Proteome (version 4), AlphaFoldDB/Swiss-Prot (version 4), CATH50, ESM Atlas High-Quality Clu30 and PDB100 require 191 GB, 3.8 GB, 3.4 GB, 1.4 GB, 110 GB and 2.0 GB RAM, respectively. The databases are kept in memory using `vmtouch` (https://github.com/hoytech/vmtouch). Databases are only required to remain resident in RAM if Foldseek is used as a webserver. During batch searches, Foldseek adapts its memory use to the available RAM of the machine. We implemented a structural visualization using the NGL viewer[44] to aid the investigation of pairwise hits. Because we only store C_α traces of the database proteins, we use PULCHRA[30] to complete the backbone of these sequences, and also of the query if necessary, to enable a ribbon visualization[45] of the proteins. For a high-quality superposition, we use TM-align[11] to superpose the structures based on the Foldseek alignment. Both PULCHRA and TM-align are executed within the users' browser using WebAssembly. They are available as `pulchra-wasm` and `tmalign-wasm` on the npm package repository as free open-source software.

Structure prediction in the webserver

We use the ESM Atlas API to predict structures of user-supplied sequences that are, at most, 400 residues long. This enables sequence-to-structure searches in the webserver.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Benchmark data are available at https://wwwuser.gwdg.de/~compbiol/foldseek.

Code availability

Foldseek is GPLv3-licensed free open-source software. The source code and binaries for Foldseek can be downloaded at https://github.com/steineggerlab/foldseek. The webserver code is available at https://github.com/soedinglab/mmseqs2-app. The analysis scripts are available at https://github.com/steineggerlab/foldseek-analysis.

References

29. Burley, S. K. et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
30. Kim, H., Mirdita, M. & Steinegger, M. Foldcomp: a library and format for compressing and indexing large protein structure sets. *Bioinformatics* **39**, btad153 (2023).
31. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).
32. Valasatava, Y. et al. Towards an efficient compression of 3D coordinates of macromolecular structures. *PLoS ONE* **12**, e0174846 (2017).
33. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
34. Farrar, M. Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**, 156–161 (2007).
35. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith–Waterman C/C++ library for use in genomic applications. *PLoS ONE* **8**, e82138 (2013).
36. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
37. Söding, J. & Remmert, M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr. Opin. Struct. Biol.* **21**, 404–411 (2011).
38. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
39. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics* **17**, 81 (2016).
40. Hung, L.-H. & Samudrala, R. fast_protein_cluster: parallel and optimized clustering of large-scale protein modeling data. *Bioinformatics* **30**, 1774–1776 (2014).
41. Jiménez-Moreno, A., Strelák, D., Filipovic, J., Carazo, J. M. & Sorzano, C. O. S. DeepAlign, a 3D alignment method based on regionalized deep learning for Cryo-EM. *J. Struct. Biol.* **213**, 107712 (2021).
42. Lafita, A. et al. BioJava 5: a community driven open-source bioinformatics library. *PLoS Comput. Biol.* **15**, e1006791 (2019).
43. Mirdita, M., Steinegger, M. & Söding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**, 2856–2858 (2019).
44. Rose, A. S. et al. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* **34**, 3755–3758 (2018).
45. Richardson, J. S. Early ribbon drawings of proteins. *Nat. Struct. Biol.* **7**, 624–625 (2000).

Acknowledgements

We thank N. Bordin, I. Sillitoe and C. Orengo for reporting issues and providing valuable feedback; Y. Zhang, P. Rotkiewicz and M. Wojdyr for making TM-align, PULCHRA and the Gemmi library freely accessible; and D.-Y. Kim for creating the Foldseek logo.

M.S. acknowledges support from the National Research Foundation of Korea (NRF) (grants 2019R1A6A1A10073437, 2020M3A9G7103933, 2021R1C1C102065 and 2021M3A9I4021220), the Samsung DS Research Fund and the Creative-Pioneering Researchers Program through Seoul National University. S.K. acknowledges support by NRF grant 2019R1A6A1A10073437. J.S. acknowledges support by the German Ministry for Education and Research (horizontal4meta). We used the compute cluster at the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG).

Author contributions

M.K., S.K., J.S. and M.S. designed the research. M.K., S.K., C.T., M.M. and M.S. developed code and performed analyses. M.K. and J.S. developed the 3Di alphabet. J.L. implemented the fast LDDT code. M.M. and C.L.M.G. developed the webserver. M.K., S.K., C.T., M.M., J.S. and M.S. wrote the manuscript.

Funding

Open access funding provided by Max Planck Society.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01773-0.

Correspondence and requests for materials should be addressed to Johannes Söding or Martin Steinegger.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

A.2 Databases

Recommended reading for Chapter 2 (listed in Section 2.11).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Bateman et al. (2024) — UniProt today; what Swiss-Prot, TrEMBL, UniParc and UniRef are, and how annotation is done.
  full text below · DOI · open access (CC BY 4.0)

  UniProt: the Universal Protein Knowledgebase in 2025. Bateman et al. (2024). *Nucleic Acids Research.* doi:10.1093/nar/gkae1010

  > The aim of the UniProt Knowledgebase (UniProtKB; https://www.uniprot.org/) is to provide users with a comprehensive, high-quality and freely accessible set of protein sequences annotated with functional information. In this publication, we describe ongoing changes to our production pipeline to limit the sequences available in UniProtKB to high-quality, non-redundant reference proteomes. We continue to manually curate the scientific literature to add the latest functional data and use machine learning techniques. We also encourage community curation to ensure key publications are not missed. We provide an update on the automatic annotation methods used by UniProtKB to predict information for unreviewed entries describing unstudied proteins. Finally, updates to the UniProt website are described, including a new tab linking protein to genomic information. In recognition of its value to the scientific community, the UniProt database has been awarded Global Core Biodata Resource status.

- Blum et al. (2024) — InterPro, the integrated classification of protein families, domains and sites.
  full text below · DOI · open access (CC BY 4.0)

  InterPro: the protein sequence classification resource in 2025. Blum et al. (2024). *Nucleic Acids Research.* doi:10.1093/nar/gkae1082

  > InterPro (https://www.ebi.ac.uk/interpro) is a freely accessible resource for the classification of protein sequences into families. It integrates predictive models, known as signatures, from multiple member databases to classify sequences into families and predict the presence of domains and significant sites. The InterPro database provides annotations for over 200 million sequences, ensuring extensive coverage of UniProtKB, the standard repository of protein sequences, and includes mappings to several other major resources, such as Gene Ontology (GO), Protein Data Bank in Europe (PDBe) and the AlphaFold Protein Structure Database. In this publication, we report on the status of InterPro (version 101.0), detailing new developments in the database, associated web interface and software. Notable updates include the increased integration of structures predicted by AlphaFold and the enhanced description of protein families using artificial intelligence. Over the past two years, more than 5000 new InterPro entries have been created. The InterPro website now offers access to 85 000 protein families and domains from its member databases and serves as a long-term archive for retired databases. InterPro data, software and tools are freely available.

- Burley et al. (2018) — the Protein Data Bank as a single global archive, and how structures are deposited and validated.
  full text below · DOI · open access (CC BY 4.0)

  Protein Data Bank: the single global archive for 3D macromolecular structure data. Burley et al. (2018). *Nucleic Acids Research.* doi:10.1093/nar/gky949

  > The Protein Data Bank (PDB) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; wwpdb.org), an international consortium that collaboratively oversees deposition, validation, biocuration, and open access dissemination of 3D macromolecular structure data. The PDB Core Archive houses 3D atomic coordinates of more than 144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules and related experimental data and metadata. Structure and experimental data/metadata are also stored in the PDB Core Archive using the readily extensible wwPDB PDBx/mmCIF master data format, which will continue to evolve as data/metadata from new experimental techniques and structure determination methods are incorporated by the wwPDB.

> Impacts of the recently developed universal wwPDB OneDep deposition/validation/biocuration system and various methods-specific wwPDB Validation Task Forces on improving the quality of structures and data housed in the PDB Core Archive are described together with current challenges and future plans.

- Bertoni et al. (2025) — the AlphaFold Protein Structure Database and how to use predicted structures.
  full text below · DOI · open access (CC BY 4.0)

  AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. Bertoni et al. (2025). *Nucleic Acids Research.* doi:10.1093/nar/gkaf1226

  > The AlphaFold Protein Structure Database (AFDB; https://alphafold.ebi.ac.uk), developed by EMBL-EBI and Google DeepMind, provides open access to hundreds of millions of high-accuracy protein structure predictions, transforming research in structural biology and the wider life sciences. Since its launch, AFDB has become a widely used bioinformatics resource, integrated into major databases, visualization platforms, and analysis pipelines. Here, we report the update of the database to align with the UniProt 2025_03 release, along with a comprehensive redesign of the entry page to enhance usability, accessibility, and structural interpretation. The new design integrates annotations directly with an interactive 3D viewer and introduces dedicated domains and summary tabs. Structural coverage has also been updated to include isoforms plus underlying multiple sequence alignments. Data are available through the website, FTP, Google Cloud, and updated APIs. Together, these advances reinforce AFDB as a sustainable resource for exploring protein sequence-structure relationships.

- Szklarczyk et al. (2024) — STRING; how functional association networks are built and scored.
  full text below · DOI · open access (CC BY 4.0)

  The STRING database in 2025: protein networks with directionality of regulation. Szklarczyk et al. (2024). *Nucleic Acids Research.* doi:10.1093/nar/gkae1113

  > Proteins cooperate, regulate and bind each other to achieve their functions. Understanding the complex network of their interactions is essential for a systems-level description of cellular processes. The STRING database compiles, scores and integrates protein-protein association information drawn from experimental assays, computational predictions and prior knowledge. Its goal is to create comprehensive and objective global networks that encompass both physical and functional interactions. Additionally, STRING provides supplementary tools such as network clustering and pathway enrichment analysis. The latest version, STRING 12.5, introduces a new 'regulatory network', for which it gathers evidence on the type and directionality of interactions using curated pathway databases and a fine-tuned language model parsing the literature. This update enables users to visualize and access three distinct network types-functional, physical and regulatory-separately, each applicable to distinct research needs. In addition, the pathway enrichment detection functionality has been updated, with better false discovery rate corrections, redundancy filtering and improved visual displays. The resource now also offers improved annotations of clustered networks and provides users with downloadable network embeddings, which facilitate the use of STRING networks in machine learning and allow cross-species transfer of protein information. The STRING database is available online at https://string-db.org/.

- The Gene Ontology Consortium et al. (2023) — the Gene Ontology knowledgebase; terms, annotations and evidence.
  full text below · DOI · open access (CC BY 4.0)

  The Gene Ontology knowledgebase in 2023. The Gene Ontology Consortium et al. (2023). *GENETICS.* doi:10.1093/genetics/iyad031

  > The Gene Ontology (GO) knowledgebase (http://geneontology.org) is a comprehensive resource concerning the functions of genes and gene products (proteins and noncoding RNAs). GO annotations cover genes from organisms across the tree of life as well as viruses, though most gene function knowledge currently derives from experiments carried out in a relatively small number of model organisms. Here, we provide an updated overview of the GO knowledgebase, as well as the efforts of the broad, international consortium of scientists that develops, maintains, and updates the GO knowledgebase. The GO knowledgebase consists of three components: (1) the GO-a computational knowledge structure describing the functional characteristics of genes; (2) GO annotations-evidence-supported statements asserting that a specific gene product has a particular functional characteristic; and (3) GO Causal Activity Models (GO-CAMs)-mechanistic models of molecular ”pathways” (GO biological processes) created by linking

multiple GO annotations using defined relations. Each of these components is continually expanded, revised, and updated in response to newly published discoveries and receives extensive QA checks, reviews, and user feedback. For each of these components, we provide a description of the current contents, recent developments to keep the knowledgebase up to date with new discoveries, and guidance on how users can best make use of the data that we provide. We conclude with future directions for the project.

The open-access papers are reproduced below under their original licences.

Nucleic Acids Research, 2025, **53**, D609–D617
https://doi.org/10.1093/nar/gkae1010
Advance access publication date: 18 November 2024
Database issue

UniProt: the Universal Protein Knowledgebase in 2025

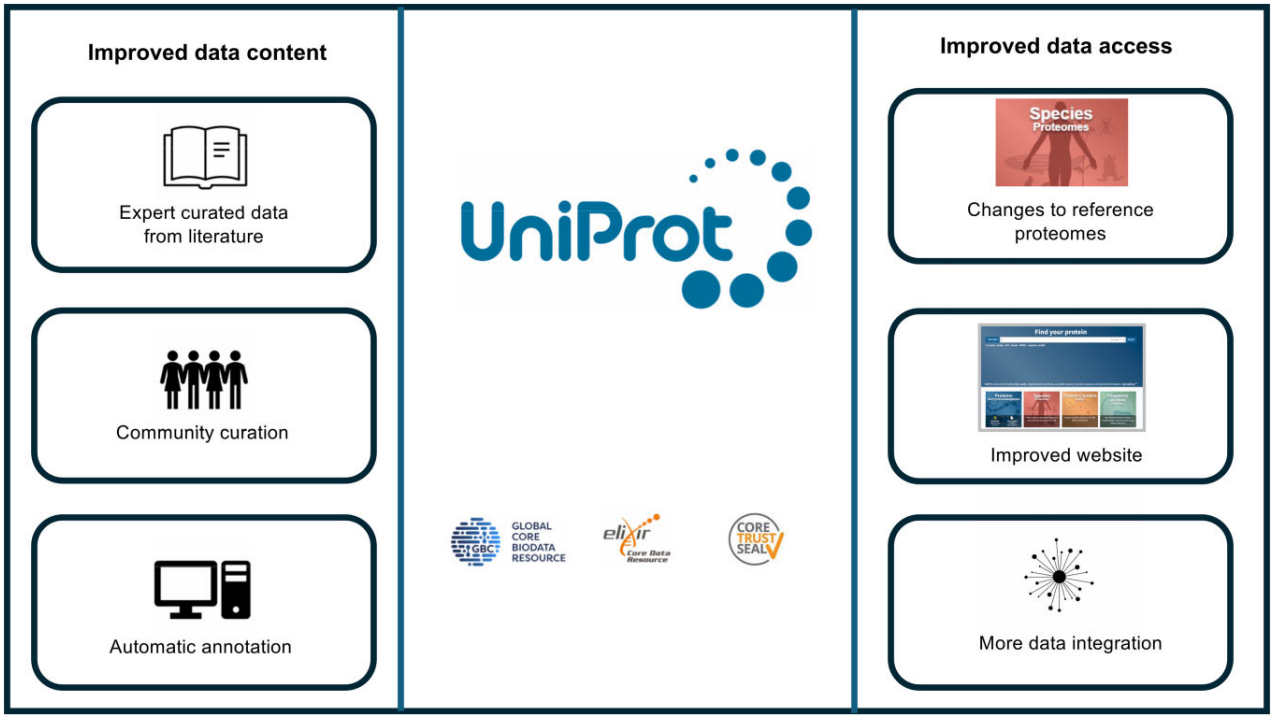
The UniProt Consortium[1,2,3,4,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK
[2]Protein Information Resource, Georgetown University Medical Center, 2115 Wisconsin Ave NW, G1 level, Suite 040A, Washington, DC 20007, USA
[3]Protein Information Resource, University of Delaware, Ammon-Pinizzotto Biopharmaceutical Innovation Building, Suite 147B, 590 Avenue 1743, Newark, DE 19713, USA
[4]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, CH-1211, Geneva 4, Switzerland
[*]To whom correspondence should be addressed. Sandra Orchard. Tel: +44 1223 494675; Email: orchard@ebi.ac.uk

Abstract

The aim of the UniProt Knowledgebase (UniProtKB; https://www.uniprot.org/) is to provide users with a comprehensive, high-quality and freely accessible set of protein sequences annotated with functional information. In this publication, we describe ongoing changes to our production pipeline to limit the sequences available in UniProtKB to high-quality, non-redundant reference proteomes. We continue to manually curate the scientific literature to add the latest functional data and use machine learning techniques. We also encourage community curation to ensure key publications are not missed. We provide an update on the automatic annotation methods used by UniProtKB to predict information for unreviewed entries describing unstudied proteins. Finally, updates to the UniProt website are described, including a new tab linking protein to genomic information. In recognition of its value to the scientific community, the UniProt database has been awarded Global Core Biodata Resource status.

Graphical abstract



Introduction

The UniProt suite of databases (https://www.uniprot.org/) serves as a leading global data resource for protein sequence and functional information (1). The long-standing critical role of the data resource in underpinning molecular biology was recognized when UniProt became one of the first databases awarded Global Core Biodata Resource status (https://globalbiodata.org) in December 2022. The cen-

Received: September 12, 2024. **Revised:** October 14, 2024. **Editorial Decision:** October 15, 2024. **Accepted:** October 16, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

tral UniProt Knowledgebase (UniProtKB) is a curated resource. It is comprised of both a reviewed set of protein entries (UniProtKB/Swiss-Prot), where each record contains a summary of the experimentally verified, or computationally predicted, functional information, added and evaluated by an expert biocurator, and the unreviewed set (UniProtKB/TrEMBL). In the latter, entries are computationally annotated by automated systems. The UniRef databases cluster sequence sets at various levels of sequence identity and the UniProt Archive (UniParc) delivers a complete set of known unique sequences, including historical obsolete sequences. If a UniParc entry sequence is not included in UniProtKB, the reason for the exclusion of that sequence is provided (e.g. pseudo-gene). In addition, each UniParc record contains a reference to its source database(s) with accession and version numbers, tagged with its status in that database, indicating if the sequence still exists or has been deleted in the source database.

Historically, UniProtKB contained all protein sequences deposited by the research community in any International Nucleotide Sequence Database Collaboration (INSDC) member resource (2) in addition to proteomes derived from selected genomes annotated by either Ensembl (3) or RefSeq (4). In the future, as the rate of whole genome sequencing increases and global projects aim to sequence across the whole biodiversity of species and, in some cases, multiple strains, sexes or variants of a single species, this will become increasingly difficult to maintain and for our users to navigate and comprehend. We describe here our response to managing the increased number of genomes now available in the public domain, and their effect on our production pipelines. In addition, we outline our work to improve the efficiency of information retrieval and summarization from the scientific literature and also from collaborating repositories, and the enhanced representation of these data in UniProtKB records.

Progress and new developments

Managing the sequence space

In May 2024, the INSDC contained 2.4 billion nucleotide sequences, many with associated translations to protein sequences made by the submitter, which form the basis of gene calls by groups such as Ensembl or RefSeq. Most of these sequences are highly redundant, i.e. identical or near identical in sequence content derived from a single organism which has been sequenced many times or from the sequencing of very closely related species. UniProtKB would not supply an optimal service to our users if we included all of the sequences available from INSDC or downstream annotation sources. Redundant sequences do not add significantly to information content, and their presence both slows down and compromises the output of services such as BLAST (5) and are costly to process and perform repetitive computation upon. UniProt release 2024_04 contains approximately 246 million sequence records in UniProtKB. This represents a relatively conservative increase in sequence number over the last 2 years (6), reflecting measures already put in place to limit the volume of data entering the Knowledgebase. As data are automatically imported from the INSDC, we filter to exclude redundant copies of proteomes from the same organism (identified by taxonomic identifier) and increasingly remove clinical isolates where, again, multiple copies of the same genome are sequenced, such as happened during the recent pandemic. Imports from Ensembl, RefSeq and other genome annotation platforms such as WormBase ParaSite (7) are still largely manually screened as import candidates. We continue to assess the quality of the remaining proteomes, using BUSCO (8) and the complete proteome detector (9). We then implement additional quality checks based on criteria developed by RefSeq, supplemented by further internally created checks. Only the highest quality, nonredundant proteomes remain to be submitted to our process for automatically selecting reference proteomes, i.e. a representative of each group of a cluster of proteomes grouped by overall sequence similarity. We additionally manually select and maintain a number of key reference proteomes, thus ensuring that those derived from model organisms or selected by the community as important for a particular area of study are tagged and remain the reference for their cluster group. To ensure stability in proteome selection, the pipeline is weighted to ensure that once selected a proteome remains representative for its group unless a related proteome which is significantly improved in quality (in terms of genome standard, coverage and annotation) becomes part of the cluster. This is then evaluated as a replacement for the existing representative proteome.

It is our intention in 2025 to limit the unreviewed sequences in UniProtKB/TrEMBL to only those derived from reference proteomes, unless there is significant additional functional information associated with an entry, for example a 3D structure, or there is a request from the community that a set of entries be included. This will initially result in a drop in the number of unreviewed entries in UniProtKB, which will then increase as biodiverse sequencing projects lead to the creation of novel proteome clusters. Redundant proteomes will continue to be available for search and download in UniProt through UniParc and we will work to improve user access to this resource. It should be noted that the selection of viral reference proteomes is a separate process. The small size of their proteome means that our redundancy pipeline is ineffective so viral reference proteomes are selected manually, in collaboration with the International Committee on Taxonomy of Viruses (10).

Expert curation

The extraction and evaluation of information relating to protein sequence, structure and function from the scientific literature and the addition of these data to the relevant UniProtKB/Swiss-Prot reviewed records is of central importance to the operation of the UniProt data resource. Expert curators target key publications and create or update records pertaining to human and model organism biology, including enzymology, host–pathogen interactions and many other areas critical to the studies of biodiversity and human health and disease. Over the last few years, a particular effort has been made to retrofit enzyme and transporter annotation using the Rhea knowledgebase of biochemical reactions, which uses the ChEBI ontology to represent reactants (11,12). At the time of writing, UniProtKB includes annotations to 12,501 Rhea reactions, which are linked to 28,259 857 UniProtKB protein sequence records, including 231,709 reviewed protein sequence records in UniProtKB/Swiss-Prot. We also actively track publications describing new enzymatic reactions or orphan reactions. One example of this is the identifi-

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

cation of the flavin reductase BLVRB (https://www.uniprot.org/uniprotkb/P30043/entry) as a moonlighting enzyme. Previously known to be involved in fetal heme catabolism in the liver, a recent publication showed that BLVRB also catalyzes the S-nitrosylation of cysteine residues of specific target proteins, including INSR and IRS1 thus inhibiting insulin signaling (13). Whilst the protein S-nitrosylation modification had been known for many years to act as a key mediator of signal transduction, and is described in more than 60 human entries in UniProt, (https://www.uniprot.org/uniprotkb?query=(keyword:KW-0702) the enzymes that mediate this modification were previously largely unknown. We curated the publication in UniProtKB/Swiss-Prot and provided functional information in the form of human readable text summaries and structured vocabularies, such as the Gene Ontology (GO) (14), Rhea or ChEBI (15). We created new reactions to describe the multistep transnitrosylation reactions in the BLVRB entry and described protein S-nitrosylation sites and their effect on target entries [HMOX2 (P30519), INSR (P06213) and IRS1 (P35568)]. The use of the ChEBI ontology to describe chemical structures for both enzyme reactions and post-translational modifications (PTMs) provides a direct link between enzymes and their targets, thereby increasing the interoperability of UniProt. UniProt curators are also highly involved in the curation of biological systems using Gene Ontology–Causal Activity Model(s) [GO-CAM(s)] (15,16) in order to describe the flow between molecular functions. We curated a GO-CAM to describe the BKVRB catalyzed reactions (https://www.alliancegenome.org/gene/HGNC:1063#function---go-annotations) and are currently working on the import and display of GO-CAM models on the UniProt website.

In a separate project, we have been proactively collaborating with research teams working in the area of antimicrobial drug resistance (AMR), focusing on mechanisms of drug resistance developed by the World Health Organization Class I target 'ESKAPE' organisms (*Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa* and *Enterobacteriaceae*). We are identifying and updating key proteins playing direct roles in AMR, such as beta-lactamases (Figure 1), carbapenemases, efflux pumps and ABC transporters are currently being targeted for annotation with efforts expanding to include members of two-component signalling systems, quorum sensing proteins and those involved in biofilm formation in the near future.

As the volume and diversity of the scientific literature increases, we are increasingly turning to machine learning (ML) techniques to assist the curation process with selection of key papers and data extraction. Existing ML frameworks such as LitSuggest (17) are used by UniProt curators to identify papers on subjects such as amino acid variants linked to human disease, enzymatic reactions and protein complexes. UniProt members have collaborated with the PubMed team at NCBI on the development of a new training and benchmarking dataset, EnzChemRed, to support the development of Natural Language Processing models to assist enzyme curation (18,19). EnzChemRED can boost the ability of BERT (Bidirectional Encoder Representations from Transformers)-based and GPT (Generative pre-trained transformer) models to extract novel enzyme–substrate relationships from the literature.

The potential use of large-language models (LLMs) to generate text directly from the scientific literature is currently being actively evaluated by UniProt curators. We are aiming to achieve an LLM-generated 'first-pass' summarization of relevant publications to add value to unreviewed entries and supplement existing information in UniProtKB/Swiss-Prot records. A number of pilot projects are currently underway evaluating the precision and recall of existing models and also the potential of LLMs to act as a curator-assist tool, increasing the throughput of publications which can be evaluated in the expert curation process.

Community curation

As previously described (6), we continue to reach out to the broader scientific community, asking researchers to alert us to entries requiring an update and novel publications carrying key pieces of information. Submitters are directed to a simple form which enables them to enter information from a publication and select appropriate annotation categories to which information can be added. A 'Batch Submission' option enables upload of a collection of publications and annotations for one or more entries. As of release 2024_04, there have been 3967 submissions enabling update of 3504 proteins with 1744 publications (https://community.uniprot.org/bbsub/STATS.html). Each of these submissions has been reviewed by a UniProt curator and submitted information is now visible on the UniProtKB webpages for each updated entry (Figure 2). Contributors are identified by their ORCID and a link to their ORCID landing page validates their expertise. We are now looking to actively solicit community submissions by identifying relevant papers through text-mining and reaching out to the authors to help evaluate the text-mining results and suggest further annotations which can be extracted from the publications. We are also exploring the use of LLMs to generate draft annotations that can be sent to paper authors for review.

Automatic annotation and ML

Adding functional information to the body of unreviewed entries in UniProtKB/TrEMBL remains an important task. These records are enriched with functional information by annotation transfer systems, which use the protein classification tool InterPro to group sequences at superfamily, family and subfamily levels, and to predict the occurrence of functional domains and important sites. The rule-based computational annotation UniRule system (20) uses these groupings to transfer experimentally verified annotations onto unstudied proteins, adding properties, such as protein name, functional annotation, catalytic activity, pathway, GO terms and subcellular location. We continue to expand this rule set, most recently with rules created by them based on NCBIfam signatures (21) evaluated and added to the set.

The Association-Rule-Based Annotator system for automatic classification and annotation of UniProtKB proteins (9) has been further developed, most recently through the incorporation of novel PANTHER signatures (22). This immediately generated 9141 new rules and 119,579,654 new predictions for over 20 million new protein sequences. We are additionally collaborating with Google Research who have developed ProtNLM, an LLM that accurately predicts descriptions of protein function directly from a protein's amino acid sequence (23). Initially, a single model was pre-trained, using the T5 framework, across the UniRef50 2018_03 dataset of ∼30 million diverse unlabelled protein sequences, then fine-tuned using labelled protein domain sequences from Pfam (24). This

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026



Figure 1. The annotation of an extended-spectrum beta-lactamase (P28585) which confers resistance to penicillins.

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026



Figure 2. Community submitted information is now added to relevant UniProtKB entries.

has since been extended to encompass the output of multiple models. To date, ProtNLM has been used to annotate more than 28 million proteins (2024_04) previously labelled as 'uncharacterized' with a functionally relevant name and we are actively exploring options with Google to extend this annotation to additional fields.

We continue to make our data more accessible to ML algorithms, to act as both training and test sets. We now make raw embeddings (per-protein and per-residue) available for UniProtKB/Swiss-Prot and some reference proteomes of model organisms. These have been generated using the bio_embeddings tool (https://github.com/sacdallago/bio_embeddings), prottrans_t5_xl_u5 model0, and can be retrieved from our Downloads page. Per-protein embeddings for UniProtKB can also be retrieved from the UniProtKB search results download function. We have also reached out to researchers to help us with our efforts to develop novel algorithms for use in the UniProtKB production pipeline and have led/participated in a number of community challenges. The UniProt metal-binding challenge invited the ML community to create computational methods to predict metal-binding sites across the whole of UniProtKB. This problem was chosen as currently ~17% of curated proteins have annotated metal-binding site residues whereas only 3% of unreviewed entries have these predicted, giving a large search space against which successful predictions could potentially be made. We provided both a training set and a test set of 1 million proteins. Unfortunately, although several groups participated in this challenge, all predictions contained a high number of false positives. More successful was our participation in a CAFA (Critical Assessment of Functional Annotation) challenge on the Kaggle platform, in which participants were asked to predict the function of a set of proteins, as indicated by the use of GO terms (www.kaggle.com/competitions/cafa-5-protein-function-prediction/). UniProt curators selected and contributed to the annotation of a training set of annotated proteins and also provided a curated test set to assist in the evaluation stage. The number of entrants was significantly higher, and a good success rate was achieved, with the possibility of embedding some of the successful code into future UniProt pipeline development.

Data integration

UniProtKB works with many other groups to access and integrate large-scale datasets, mapping these data onto the appropriate protein sequence records and displaying the mappings via the ProtVista Feature Viewer (25). All data shown are downloadable via File Transfer Protocol and Application Programming Interfaces (APIs). We have worked to improve our support for the mass spectrometry (MS) community (26). The existing UniProt MS peptide identification pipeline by which MS-based proteomic data deposited into proteomeXchange databases (27,28) is reprocessed, scored and used to verify the existence of each protein is being upgraded to meet the stringency of the HUPO Human Proteome Project (HPP) 3.0 guidelines (29). UniProtKB now actively supports the HUPO HPP which aims to provide MS evidence for the existence of each protein in the human proteome (Omenn *et al.* 2024, in preparation). All proteins identified by the qualifying number of unique peptides of a set minimum length are annotated with the keyword 'proteomics identification', and the protein existence level is set to 'experimental evidence at protein level'.

We have additionally collaborated with PRIDE (29), PeptideAtlas (27,28) and the University of Liverpool to enable the integration of PTMs from these resources into UniProtKB records (30). Filtering and reanalysis ensure that only high-quality datasets are imported and reanalysed, with these datasets stored in PRIDE attributed a unique ID (PXD), thereby facilitating user access, traceability and reusability. Modified sites are assigned a confidence score based on their false localization rate across multiple datasets to reflect the strength of evidence available. Data are integrated into UniProtKB and visualized in the ProtVista Feature Viewer

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

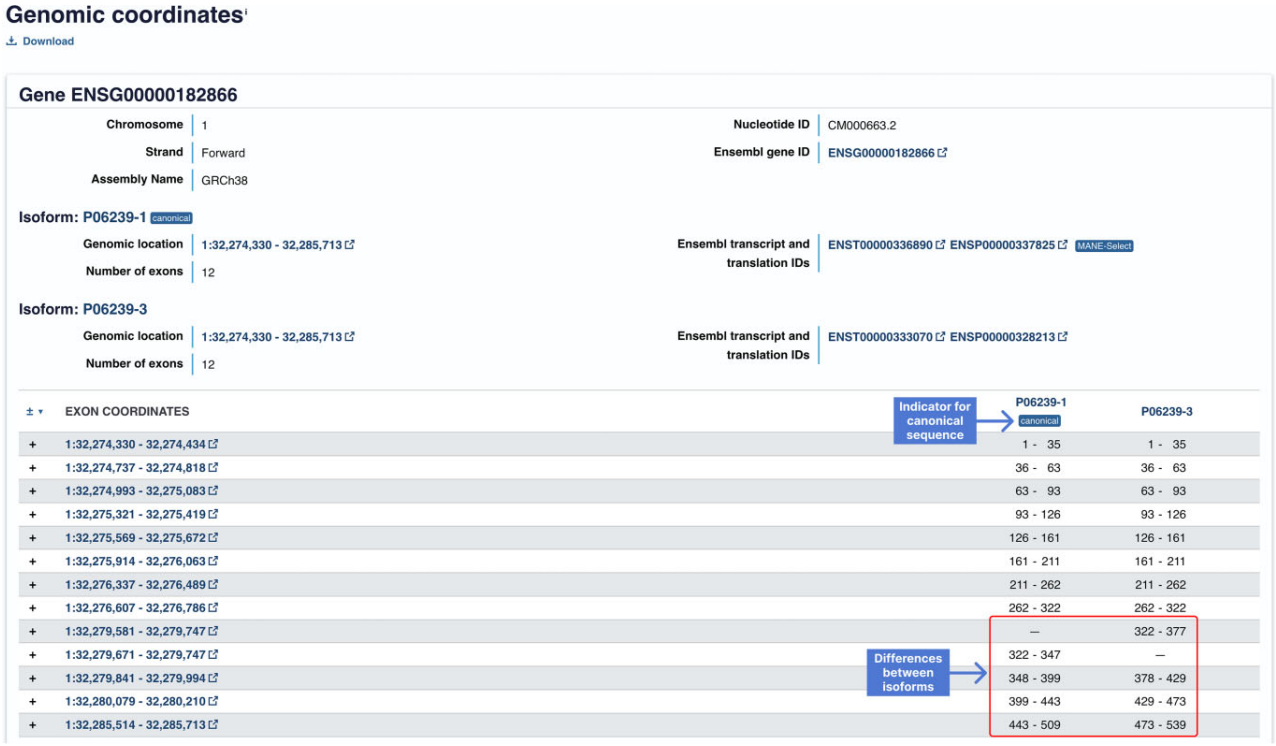


Figure 3. The Genomics Tab view for the human Tyrosine-protein kinase Lck protein (P06239).

both in the PTM/processing section of the protein entry page and in the Feature Viewer in site-centric format. Modified peptide data are visualized in the proteomics section of the Feature Viewer, and both site-centric and peptide-centric data are accessible via the proteins API. At the time of writing, rice (*Oryza sativa subsp. Japonica)* and *Plasmodium falciparum* proteomic PTM data are available with *Saccharomyces cerevisiae*, mouse (*Mus musculus*) and human (*Homo sapiens*) data currently being processed.

Website

We continue to develop the UniProt website with an emphasis on providing the user with an effective search, easy navigation and excellent responsiveness whilst further improving the tools dashboard and the programmatic access interface (API). The website has been built using a modular approach, separating the front end (web user interface) from the back end (API). We have added additional views to the website, first moving the variant view from its previous position embedded in the disease and variant section of the website to a separate tab to increase the overall speed of website loading. Second, we have created a genomics tab (Figure 3) which enables the user to link protein data to the corresponding genomic coordinates. Genomic coordinates of UniProtKB proteins are imported from the corresponding nucleotide sequence data in the INSDC. For those imported from Ensembl, Ensembl Genome, RefSeq and WormBase ParaSite, the genomic coordinates are also accessed from those resources. For each UniProtKB entry, representing the product(s) of one gene, details of the genomic assembly, chromosomal location of the gene and strand are given. Each protein isoform is listed with its corresponding genomic location, the total number and the genomic coordinates of the exons creating each isoform and the corresponding ranges of encoded amino acids.

Visualizing proteins in the context of protein complexes

Many proteins are functionally operative only as a member of a protein complex, a set of polypeptide chains, that interact with each other and/or with a nucleic acid. Mutations in genes encoding different proteins within the same complex cause similar diseases (locus heterogeneity). In order to present the user with a representation of a protein in its functional environment, we have collaborated with the Complex Portal (31) to provide structured, machine-readable descriptions of complexes using UniProtKB identifiers and a range of shared reference ontologies, such as the GO and ChEBI. To date, a total of almost 5000 manually curated complexes have been released with UniProtKB curators contributing significantly to that number. Complex Portal entries are cross-referenced in UniProtKB and we have added the ComplexViewer (32) to the UniProt website. This provides a dynamic, interactive, responsive 2D visualization of each complex, enabling users to see its topology and stoichiometry (when known) and zoom in on binding regions down to the residue level (Figure 4). Interactive links enable users to easily move between the UniProtKB entries describing complex components or to learn more about the full assembly in the Complex Portal.

Conclusion

The UniProt database will undergo some fundamental changes over the next 1–2 years, as we adapt our data process-

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

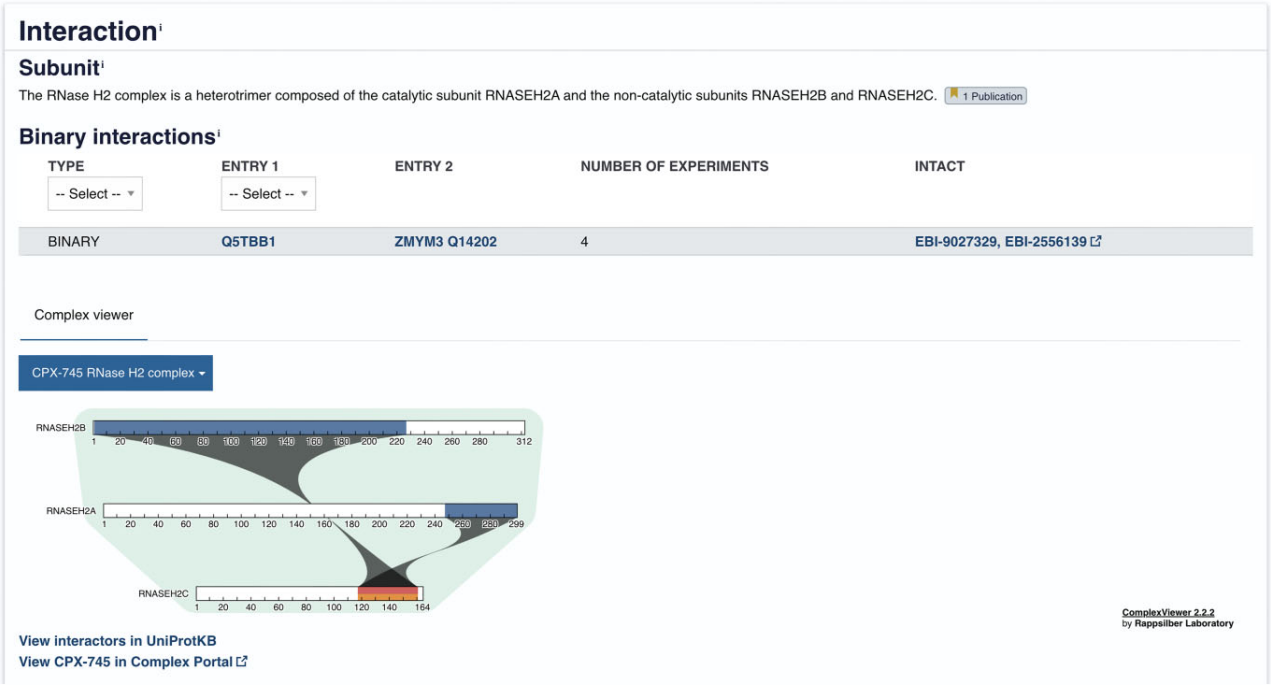


Figure 4. Visualization of the RNase H2 complex as seen in the UniProt entry for the Ribonuclease H2 subunit A protein (O75792).

ing pipelines to address the increase in the amount of whole genome sequencing data which is now being produced on an almost industrial scale. At the same time, we wish to ensure our users enjoy the same quality of service we have offered them over the previous decades. We aim to present a reference proteome for each taxonomic grouping to the research community. Whilst we will continue to evaluate the scientific literature and summarize key information on proteins and their functions, new ML tools and language models mean we are approaching how we do so in a very different way. However, we will ensure that we maintain the high-quality, expert curation for which UniProtKB is known. We additionally encourage community contributions and now make those visible directly on the webpage of the relevant entries. We will continue to develop our methods for the computational annotation of unreviewed proteins and will increasingly work with the ML community through collaborations and challenges to enable this.

UniProtKB acts as a central hub of information integrating data from many external resources. We supply a number of APIs enabling users to access UniProt data or incorporate this into their own resources (www.uniprot.org/help/programmatic_access). We proactively collaborate with other data producers and data (re)users and are always looking for opportunities to extend such interactions. We greatly value the feedback and annotation updates from our user community. Please send your comments and suggestions via the contact link on the UniProt website (https://www.uniprot.org/contact).

Data availability

UniProt releases are published every 8 weeks. We provide customizable views and downloads in a range of formats via the website, and file sets at the FTP site (www.uniprot.org/downloads), and supply users with a number of different options for computational access to the data (www.uniprot.org/help/programmatic_access). These include the website RESTful Application Programming Interface (API), stable URLs that can be bookmarked, linked and reused, the SPARQL API that allows users to perform complex queries across all UniProt data and also other resources that provide a SPARQL endpoint and a Java API.

Funding

National Human Genome Research Institute (NHGRI) [HG002273]; Office of the Director, NIH OD [U24HG007822]; National Institute of Allergy and Infectious Diseases (NIAID); National Institute on Aging (NIA); National Institute of General Medical Sciences (NIGMS) [R35GM141873]; National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); National Eye Institute (NEI); National Cancer Institute (NCI); National Heart, Lung, and Blood Institute (NHLBI); Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T015608/1]; National Science Foundation's Directorate for Biological Sciences [BB/X002179/1]; Open Targets; Horizon 2020 - Research and Innovation Framework Programme [825575]; State Secretariat for Education, Research and Innovation (SERI); European Molecular Biology Laboratory (EMBL) Australia. Funding for open access charge: National Institutes of Health.

Conflict of interest statement

None declared.

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

References

1. Lussi,Y.C., Magrane,M., Martin,M.J., Orchard,S. and UniProt Consortium (2023) Searching and navigating UniProt databases. *Curr. Protoc.*, **3**, e700.
2. Arita,M., Karsch-Mizrachi,I. and Cochrane,G. (2021) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **49**, D121–D124.
3. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O., Azov,A.G., Barba,M., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2024) Ensembl 2024. *Nucleic Acids Res.*, **52**, D891–D899.
4. Sayers,E.W., Beck,J., Bolton,E.E., Brister,J.R., Chan,J., Comeau,D.C., Connor,R., DiCuccio,M., Farrell,C.M., Feldgarden,M., *et al.* (2024) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **52**, D33–D43.
5. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
6. UnitProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
7. Howe,K.L., Bolt,B.J., Shafie,M., Kersey,P. and Berriman,M. (2017) WormBase ParaSite—a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.*, **215**, 2–10.
8. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
9. UniProt Consortium (2021) UniProt: the Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
10. Lefkowitz,E.J., Dempsey,D.M., Hendrickson,R.C., Orton,R.J., Siddell,S.G. and Smith,D.B. (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.*, **46**, D708–D717.
11. Morgat,A., Lombardot,T., Coudert,E., Axelsen,K., Neto,T.B., Gehant,S., Bansal,P., Bolleman,J., Gasteiger,E., de Castro,E., *et al.* (2020) Enzyme annotation in UniProtKB using Rhea. *Bioinformatics*, **36**, 1896–1901.
12. Bansal,P., Morgat,A., Axelsen,K.B., Muthukrishnan,V., Coudert,E., Aimo,L., Hyka-Nouspikel,N., Gasteiger,E., Kerhornou,A., Neto,T.B., *et al.* (2022) Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.*, **50**, D693–D700.
13. Zhou,H.-L., Grimmett,Z.W., Venetos,N.M., Stomberski,C.T., Qian,Z., McLaughlin,P.J., Bansal,P.K., Zhang,R., Reynolds,J.D., Premont,R.T., *et al.* (2023) An enzyme that selectively S-nitrosylates proteins to regulate insulin signaling. *Cell*, **186**, 5812–5825.
14. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
15. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
16. Thomas,P.D., Hill,D.P., Mi,H., Osumi-Sutherland,D., Van Auken,K., Carbon,S., Balhoff,J.P., Albou,L.-P., Good,B., Gaudet,P., *et al.* (2019) Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.*, **51**, 1429–1433.
17. Allot,A., Lee,K., Chen,Q., Luo,L. and Lu,Z. (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, **49**, W352–W358.
18. Lai,P.-T., Coudert,E., Aimo,L., Axelsen,K., Breuza,L., de Castro,E., Feuermann,M., Morgat,A., Pourcel,L., Pedruzzi,I., *et al.* (2024) EnzChemRED, a rich enzyme chemistry relation extraction dataset. *Scientific Data*, **11**, 982.
19. Lai,P.-T., Coudert,E., Aimo,L., Axelsen,K., Breuza,L., de Castro,E., Feuermann,M., Morgat,A., Pourcel,L., Pedruzzi,I., *et al.* (2024) EnzChemRED, a rich enzyme chemistry relation extraction dataset [Data set]. Zenodo, https://zenodo.org/records/11067998.
20. MacDougall,A., Volynkin,V., Saidi,R., Poggioli,D., Zellner,H., Hatton-Ellis,E., Joshi,V., O'Donovan,C., Orchard,S., Auchincloss,A.H., *et al.* (2020) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics*, **36**, 4643–4648.
21. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R., *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
22. Thomas,P.D., Ebert,D., Muruganujan,A., Mushayahama,T., Albou,L.-P. and Mi,H. (2021) PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.*, **31**, 8–22.
23. Paysan-Lafosse,T., Blum,M., Chuguransky,S. and Grego,T. (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
24. Brevdo,E. (2023) ProtNLM: Model-based Natural Language Protein Annotation. https://ebrevdo.github.io.
25. Salazar,G.A., Luciani,A., Watkins,X., Kandasaamy,S., Rice,D.L., Blum,M., Bateman,A. and Martin,M. (2023) Nightingale: web components for protein feature visualization. *Bioinform. Adv.*, **3**, vbad064.
26. Bowler-Barnett,E.H., Fan,J., Luo,J., Magrane,M., Martin,M.J., Orchard,S. and UniProt Consortium (2023) UniProt and mass spectrometry-based proteomics—a 2-way working relationship. *Mol. Cell. Proteomics*, **22**, 100591.
27. Deutsch,E.W., Lane,L., Overall,C.M., Bandeira,N., Baker,M.S., Pineau,C., Moritz,R.L., Corrales,F., Orchard,S., Van Eyk,J.E., *et al.* (2019) Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J. Proteome Res.*, **18**, 4108–4116.
28. Deutsch,E.W., Bandeira,N., Perez-Riverol,Y., Sharma,V., Carver,J.J., Mendoza,L., Kundu,D.J., Wang,S., Bandla,C., Kamatchinathan,S., *et al.* (2023) The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res.*, **51**, D1539–D1548.
29. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M., *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, **50**, D543–D552.
30. Kalyuzhnyy,A., Eyers,P.A., Eyers,C.E., Bowler-Barnett,E., Martin,M.J., Sun,Z., Deutsch,E.W. and Jones,A.R. (2022) Profiling the human phosphoproteome to estimate the true extent of protein phosphorylation. *J. Proteome Res.*, **21**, 1510–1524.
31. Meldal,B.H.M., Perfetto,L., Combe,C., Lubiana,T., Ferreira Cavalcante,J.V., Bye-A-Jee,H., Waagmeester,A., Del-Toro,N., Shrivastava,A., Barrera,E., *et al.* (2023) Complex Portal 2022: new curation frontiers. *Nucleic Acids Res.*, **50**, D578–D586.
32. Combe,C.W., Sivade,M.D., Hermjakob,H., Heimbach,J., Meldal,B.H.M., Micklem,G., Orchard,S. and Rappsilber,J. (2017) ComplexViewer: visualization of curated macromolecular complexes. *Bioinformatics*, **33**, 3673–3675.

Appendix

The UniProt Consortium

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H. Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L. Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu and Rossana Zaru at the EMBL-European

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Bioinformatics Institute; Alan J. Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H. Auchincloss, Kristian B. Axelsen, Parit Bansal, Delphine Baratin, Teresa M. Batista Neto, Marie-Claude Blatter, Jerven T. Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L. Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J. A. Sigrist, Karin Sonesson, Shyamala Sundaram and Anastasia Sveshnikova at the SIB Swiss Institute of Bioinformatics; Cathy H. Wu, Cecilia N. Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Yuqi Wang and Jian Zhang at the Protein Information Resource.

Downloaded from academic.oup.com/nar/article/53/D1/D609/7902999 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Received: September 12, 2024. **Revised:** October 14, 2024. **Editorial Decision:** October 15, 2024. **Accepted:** October 16, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Nucleic Acids Research, 2025, **53**, D444–D456
https://doi.org/10.1093/nar/gkae1082
Advance access publication date: 20 November 2024
Database issue

InterPro: the protein sequence classification resource in 2025

Matthias Blum [1,*], Antonina Andreeva [1], Laise Cavalcanti Florentino [1], Sara Rocio Chuguransky [1], Tiago Grego [1], Emma Hobbs [1], Beatriz Lazaro Pinto [1], Ailsa Orr [1], Typhaine Paysan-Lafosse [1], Irina Ponamareva [1], Gustavo A. Salazar [1], Nicola Bordin [2], Peer Bork [3], Alan Bridge [4], Lucy Colwell [5], Julian Gough [6], Daniel H. Haft [7], Ivica Letunic [8], Felipe Llinares-López [9], Aron Marchler-Bauer [7], Laetitia Meng-Papaxanthos [10], Huaiyu Mi [11], Darren A. Natale [12], Christine A. Orengo [2], Arun P. Pandurangan [6], Damiano Piovesan [13], Catherine Rivoire [4], Christian J.A. Sigrist [4], Narmada Thanki [7], Françoise Thibaud-Nissen [7], Paul D. Thomas [11], Silvio C.E. Tosatto [13,14], Cathy H. Wu [12] and Alex Bateman [1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK
[2]Department of Structural and Molecular Biology, University College London, Gower St, Bloomsbury, London WC1E 6BT, UK
[3]European Molecular Biology Laboratory, Structural and Computational Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany
[4]Swiss-Prot Group, Swiss Institute of Bioinformatics, CMU, 1 rue Michel Servet, CH-1211, Geneva, Switzerland
[5]Google DeepMind, Cambridge, MA 02142, USA
[6]Medical Research Council Laboratory of Molecular Biology, Cambridge Biomedical Campus, Francis Crick Ave, Trumpington, Cambridge CB2 0QH, UK
[7]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA
[8]Biobyte Solutions GmbH, Bothestr 142, 69126 Heidelberg, Germany
[9]Google DeepMind, 75009 Paris, France
[10]Google DeepMind, 8002 Zürich, Switzerland
[11]Division of Bioinformatics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA 90033, USA
[12]Protein Information Resource, Georgetown University Medical Center, WA, DC 20007, USA
[13]Department of Biomedical Sciences, University of Padova, Padova 35121, Italy
[14]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari 70126, Italy

[*]To whom correspondence should be addressed. Email: mblum@ebi.ac.uk

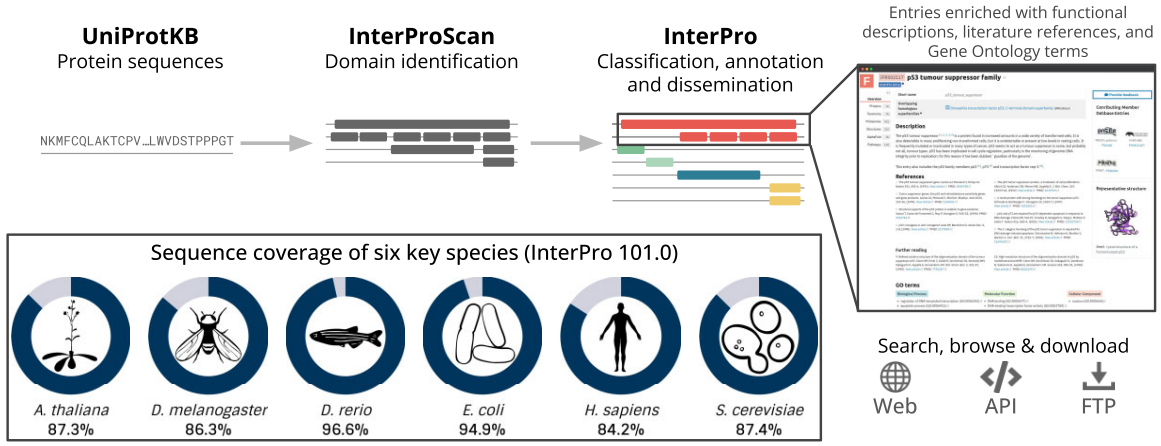
Abstract

InterPro (https://www.ebi.ac.uk/interpro) is a freely accessible resource for the classification of protein sequences into families. It integrates predictive models, known as signatures, from multiple member databases to classify sequences into families and predict the presence of domains and significant sites. The InterPro database provides annotations for over 200 million sequences, ensuring extensive coverage of UniProtKB, the standard repository of protein sequences, and includes mappings to several other major resources, such as Gene Ontology (GO), Protein Data Bank in Europe (PDBe) and the AlphaFold Protein Structure Database. In this publication, we report on the status of InterPro (version 101.0), detailing new developments in the database, associated web interface and software. Notable updates include the increased integration of structures predicted by AlphaFold and the enhanced description of protein families using artificial intelligence. Over the past two years, more than 5000 new InterPro entries have been created. The InterPro website now offers access to 85 000 protein families and domains from its member databases and serves as a long-term archive for retired databases. InterPro data, software and tools are freely available.

Received: September 13, 2024. **Revised:** October 11, 2024. **Editorial Decision:** October 14, 2024. **Accepted:** October 23, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Graphical abstract



Introduction

The rapid advancements in genomic technologies, coupled with substantial decreases in the cost of sequencing, have brought forth an era where large-scale sequencing projects generate data at an unprecedented rate. However, the surge in data volume has outpaced our capabilities to characterise sequences through experimental methods alone. Consequently, there is a strong need for sophisticated computational tools and databases to manage, annotate and interpret these vast repositories of biological information.

Launched in 1999, InterPro aims to meet this challenge by integrating diverse information about protein families, domains and functional sites from specialist member databases into a single cohesive resource. These databases include CATH-Gene3D (1), the Conserved Domains Database (CDD) (2), HAMAP (3), NCBIFAM (4), PANTHER (5), PIRSF (6), Pfam (7), PRINTS (8), PROSITE (9), the Structure-Function Linkage Database (SFLD) (10), SMART (11) and SUPERFAMILY (12). They utilise predictive models – such as profile hidden Markov models (HMM), position-specific scoring matrices, and regular expressions, collectively referred to as *signatures* – to search protein sequences and assign potential functions.

InterPro consolidates and cross-references these diverse annotations, producing a comprehensive overview of protein families, domains and functional sites. Each InterPro entry is assigned a unique accession number, name and short name and is annotated with a descriptive abstract and Gene Ontology (GO) terms (13) that can be assigned to all sequences matched by that entry (14). This integrative approach not only reduces redundancy but also enhances the robustness of protein annotations available to the scientific community.

InterPro also provides comprehensive information about sequence features, including consensus annotations of long-range intrinsic disorder regions predicted by MobiDB-lite (15), and predictions of signal peptides, transmembrane regions and coiled-coils using the SignalP (16), Phobius (17), TMHMM (18) and Coils (19) software packages. Additionally, InterPro offers functional classification of protein domains through FunFam (20), facilitating the identification of evolutionary relationships and functional properties. Furthermore, InterPro includes annotations from AntiFam (21), a tool designed to detect spurious open reading frames (ORFs).

InterPro and its associated software are extensively disseminated and utilised by the scientific community, such that it is recognised as a Core Data Resource by ELIXIR (22) and as a Global Core Biodata Resource by the Global Biodata Coalition (23). Over 40 million sequence searches were processed between January 2023 and June 2024 by the InterProScan web service, which facilitates the analysis of user-submitted sequences via the EMBL-EBI Job Dispatcher (24).

A primary application of InterPro data is its integration into UniProtKB (25), where InterPro annotations underpin the automatic annotation of proteins (26). To support this task, InterPro matches are computed weekly using the InterProScan software package (27) against new sequences deposited in the UniProt Archive (UniParc), a comprehensive and non-redundant database containing the majority of publicly available protein sequences. This process ensures that new sequences are promptly and accurately annotated. Additionally, InterPro data are integral to various annotation pipelines, including Ensembl (28), Ensembl Genomes (29), PDBe (30), Blast2GO (31), Genome Properties (32), MGnify (33) and MGnify Genomes (34).

InterPro's release cycle aligns with UniProt, making new releases of InterPro publicly available every 8 weeks, 1 day after each UniProt release.

Data update

Member database signatures are not automatically incorporated into InterPro; instead, they undergo a thorough manual inspection and integration process. Matches between these signatures and the most recent version of UniProtKB are evaluated to eliminate any false positives. Signatures corresponding to the same biological entity are consolidated into unified InterPro entries, thereby reducing redundancy. For instance, signatures from CDD, PROSITE Profiles, Pfam and SMART, which represent the CUB domain (cd00041, PS01180, PF00431, SM00042, respectively), are integrated into a single InterPro entry (IPR000859).

New InterPro entries are subject to manual annotation, including assignment of a name, short name, descriptive abstract and GO terms that are applied consistently to all matched proteins. Hierarchical relationships are established among evolutionarily related InterPro entries, linking smaller, functionally specific subfamilies of larger families, or subclasses of broader

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Table 1. Release version and number of member database signatures integrated into InterPro version 101.0

Member database	Release number	Total signatures	Integrated signatures
CATH-Gene3D	4.3.0	6631	2822 (42.6%)
CDD	3.20	18 882	4908 (26.0%)
HAMAP	2023_05	2389	2384 (99.8%)
NCBIFAM	15.0	7289	5884 (80.7%)
PANTHER	18.0	15 693	10 520 (67.0%)
Pfam	37.0	21 979	21 171 (96.3%)
PIRSF	3.10	3285	3230 (98.3%)
PRINTS	42.0	2106	1939 (92.1%)
PROSITE Patterns	2023_05	1311	1283 (97.9%)
PROSITE Profiles	2023_05	1379	1,21 (93.6%)
SFLD	4	303	159 (52.5%)
SMART	9.0	1322	1274 (96.4%)
SUPERFAMILY	1.75	2019	1643 (81.4%)

Table 2. Coverage of UniProtKB and UniParc (non-redundant archive of protein sequences) by InterPro entries (version 101.0)

Sequence database	Number of sequences	Number of sequences with one or more matches to InterPro
UniProtKB/reviewed	571 864	553 706 (96.8%)
UniProtKB/unreviewed	245 324 902	200 675 244 (81.8%)
UniProtKB (total)	245 896 766	201 228 950 (81.8%)
UniParc	808 031 558	654 323 531 (81.0%)

classes of domains. Bi-monthly reviews are conducted on annotation and sequence match information following UniProtKB updates, with InterPro entries adjusted based on revised sequence information or updated biological knowledge (e.g. newly characterised protein functions). Despite requiring significant curation effort, these updates are crucial for maintaining annotation accuracy, especially given the dynamic nature of the underlying data.

With each release of InterPro, one or more of its constituent member databases may undergo updates, resulting in the continuous addition of new signatures and entries into the resource. Since the last update publication (35), there have been 11 public releases of InterPro, during which 8734 new signatures have been integrated into 7963 new InterPro entries. The most recent release, version 101.0, includes 84 588 member database signatures, 69% of which are integrated into 45 899 InterPro entries. The percentage of signatures integrated into InterPro for each member database is shown in Table 1.

In UniProtKB, 81.8% of the sequences have at least one match to an InterPro entry (see Table 2). Although sequence coverage has decreased by 0.2% over the past 2 years, this decline should be considered in the context of substantial growth within UniProtKB during the same period, with InterPro annotations now covering an additional 18 million sequences. Meanwhile, the number of sequences in UniParc has increased by 56%, and InterPro coverage of UniParc has grown by 1.1%.

As UniProtKB continues to expand, maintaining or improving InterPro coverage is increasingly challenging. Larger, well-characterised protein families have already been integrated, and newer entries often focus on more narrowly defined taxonomic groups. Despite a 371% increase in the number of sequences in UniProtKB over the past decade, InterPro has managed to sustain a sequence coverage above 80% (see Figure 1A). This highlights ongoing efforts to enhance annotation across a rapidly growing sequence database.

Beyond sequence coverage, we have also evaluated the amino acid residue coverage of InterPro and its member databases. Overall, 89.8% of residues in UniProtKB receive some level of annotation. Figure 1B illustrates the cumulative unique residue coverage: (i) InterPro entries cover 64.4 billion out of 87 billion residues (74%); (ii) signatures from member databases pending integration into InterPro cover 4.2%; (iii) residues found in intrinsically disordered regions account for 3.3%; (iv) residues associated with other sequence features, such as coiled-coil regions, transmembrane regions and signal peptides, cover 8.3%; and (v) residues with no annotation make up the remaining portion.

New developments

InterPro: the new home of Pfam data

As reported in our previous publication, we decided to decommission the Pfam website (formerly accessible at http://pfam.xfam.org) due to challenges in maintaining and updating its legacy infrastructure. In August 2022, we posted a notice on the Pfam website, informing users of an impending automatic redirection to the InterPro website starting in October 2022. With the initiation of this redirection in October 2022, we provided access to a legacy version of the Pfam website at http://pfam-legacy.xfam.org for users who had not yet migrated.

Although the legacy site was initially scheduled for shutdown in January 2023, it remained available until Summer 2023 to facilitate user transition. This transition was supported by several blog posts, X (formerly Twitter) updates and webinars to ensure a smooth migration of the Pfam userbase to InterPro. The former Pfam website (http://pfam.xfam.org) now hosts a static page that replicates the appearance of the old site, with its features (search, browse, etc.) redirecting to the corresponding pages on the InterPro website.

Pfam is now exclusively accessible through InterPro, where all critical features from the original Pfam site are available. Users can also access Pfam data programmatically via the InterPro REST API and annotate sequences with Pfam through InterProScan, which is available both as a web service and a downloadable standalone program. By moving away from a legacy codebase with increasing maintenance demands, Pfam now focuses more on its core mission of building protein families. This shift allows for more frequent updates, which will be seamlessly integrated into InterPro releases. All past Pfam release files remain accessible via the EMBL-EBI FTP site, with future release files to be made available there as well.

Leveraging AI to describe protein families

InterPro entries are derived from member database signatures, with annotations for new entries often sourced from these databases. When member database signatures lack annotations, Interpro curators must create descriptions *de novo*, frequently using scientific literature. This process is laborious and slows down the integration of signatures. This issue is particularly notable for PANTHER, a vast collection of protein families. PANTHER ranks as the third largest member database by the number of signatures it provides (preceded by CDD and Pfam) and as the second largest by the number of UniProtKB sequences matched (after Pfam). However, no PANTHER signatures are accompanied by descriptive abstracts, resulting in

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

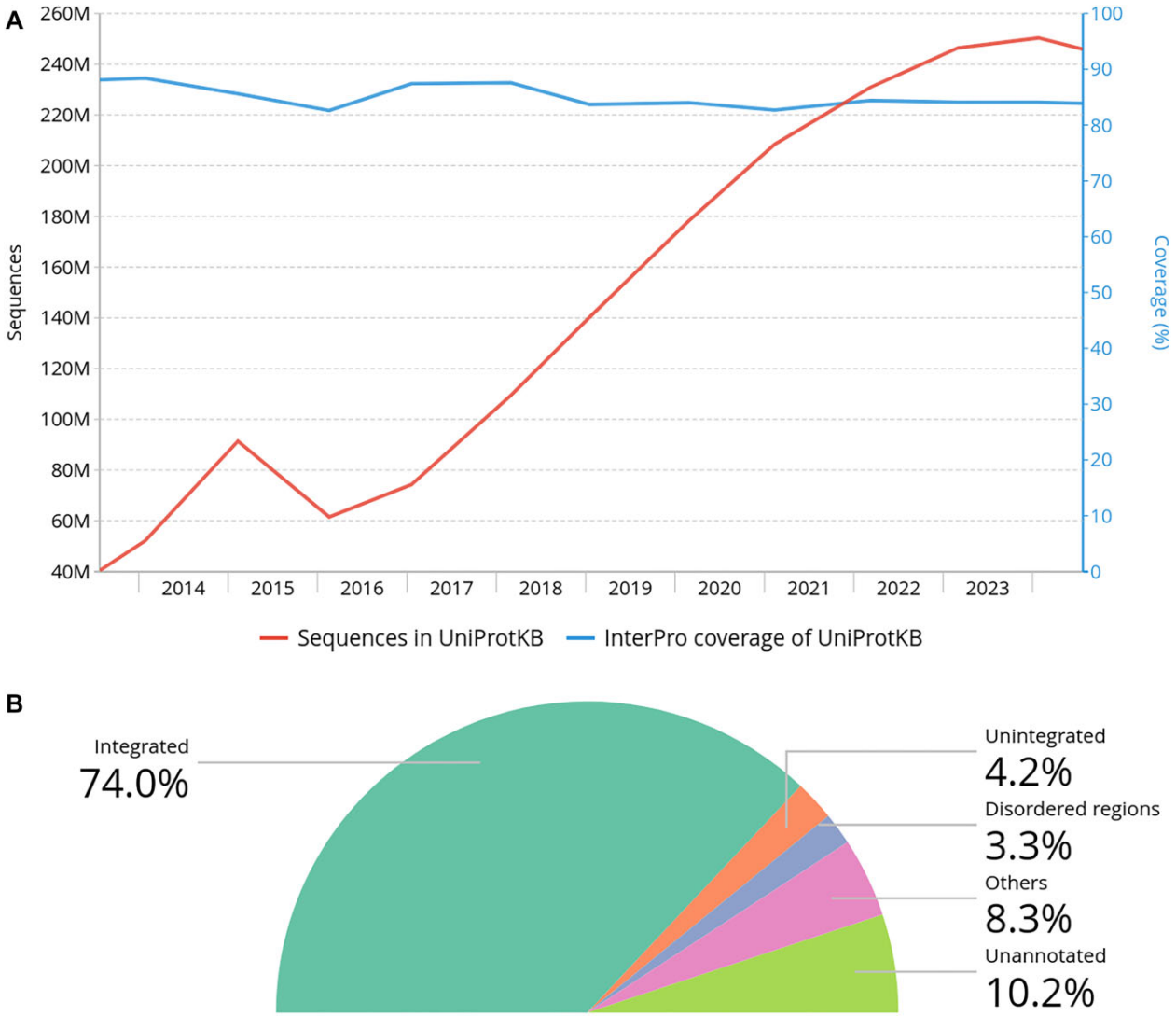


Figure 1. InterPro coverage of UniProtKB. (**A**) InterPro coverage of UniProtKB sequences alongside the growth of UniProtKB over time (January 2014–July 2024). (**B**) InterPro coverage of amino acid residues in UniProtKB, categorised as follows: residues covered by signatures already integrated into InterPro, signatures from member databases that are awaiting integration, intrinsically disordered regions and regions predicted to be signal peptides, transmembrane domains, or coiled-coils. The remaining residues are classified as unannotated.

a lower integration rate. Moreover, the rising publication rate, combined with a limited number of InterPro curators, exacerbates the challenge.

Recent advancements in large language models (LLMs) have reached a quality level sufficient for application in some curation tasks. LLMs, a class of machine learning models featuring billions of parameters, excel at predicting subsequent tokens in a sequence. A notable application of LLMs is LitSumm (36), a method that generates automated summaries for non-coding RNA (ncRNA) genes. LitSumm extracts and selects relevant sentences from the literature using ncRNA identifiers and names, instructs a LLM to generate a summary, and then further directs the LLM to check and refine the summary for consistency. However, this method is less effective with PANTHER signatures since their accession numbers are infrequently cited in publications and their names are often non-descriptive (e.g. Uncharacterised, Expressed protein, Hypothetical protein). Nonetheless, 70% of unintegrated PANTHER signatures match at least one sequence from Swiss-Prot, the manually curated section of UniProtKB. To address the lack of descriptions, we developed an alternative approach using GPT-4 models from OpenAI (37). For PANTHER signatures with Swiss-Prot matches, we extracted relevant information from Swiss-Prot entries, including protein names and general annotations (function, activity regulation, similarity). We then instructed the model to summarise the overall function of the family using this information and the knowledge embedded in the model. Manual assessments of a subset of these summaries showed high-quality ratings for the majority of them. Consequently, we released 5491 descriptions of PANTHER families in InterPro 97.0. We further refined our method to include the generation of names and short names for each family, facilitating the automatic creation of InterPro entries.

Between InterPro 99.0 and 101.0, we automatically generated 3887 InterPro entries based on AI-generated annotations for 3316 PANTHER families, 507 NCBIFAM families, and 64 CATH-Gene3D superfamilies. These entries are clearly

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

marked as AI-generated and unreviewed on the InterPro website (see Figure 2). Each page containing AI-generated annotations includes a warning notice with a link to the InterPro documentation, explaining how the annotations were generated, which model was used, and providing an example input context. This documentation will be updated as needed if we change the annotation method or model. Additionally, we updated our internal curation application to allow curators to review and approve these annotations, with the reviewed status reflected in subsequent releases.

Employing smaller local models is one of the possibilities of speeding up and decreasing the cost of using LLMs. We explored this possibility with fine-tuning a SciPhi-Mistral-7B-32k model using PANTHER families, by learning to imitate GPT-4 responses. This particular model was pretrained from Mistral 7B v0.1 (38) on scientific texts, which makes it suitable for describing protein families.

To fine-tune the SciPhi-Mistral model, we used a collection of 5559 PANTHER families that we had previously described using GPT-4. For each family, we provided a prompt containing the information from Swiss-Prot proteins and specific instructions for GPT-4, along with the corresponding response generated by GPT-4 based on the prompt. The only change in the prompts was additional markdown for the SciPhi-Mistral model. The markdown contained tags [SYS], [CONT], [INST], [ANS] (for 'system', 'context', 'instruction' and 'answer', respectively) and corresponding closing tags helping the model to parse the input. The dataset was divided into training (5473 families) and testing (86 families) subsets. We ran the fine-tuning for one epoch, where the SciPhi-Mistral model was learning to generate responses similar to GPT-4 responses.

To run the evaluation, we randomly selected 56 families from the testing subset and generated three types of responses for each: (i) a response generated by GPT-4; (ii) a response generated by the baseline, non-fine-tuned SciPhi-Mistral-7B-32k model; (iii) a response generated by our fine-tuned SciPhi-Mistral-7B-32k model.

To compare the models, we created 168 response pairs by considering all possible pairwise comparisons for each of the 56 families. These pairs were loaded into Label Studio, where a labelling task was created for each of the 5 human InterPro curators. For each pair, we asked the curators to decide which of the two models' responses they prefer. No other information except for the textual descriptions was given. The evaluation was performed independently by each curator. After this, their responses were aggregated as votes with the same weight. The results of the evaluation are presented in Figure 3. The baseline model predictions are outperformed by both GPT-4 and the fine tuned model by a large margin – GPT-4 was preferred to the baseline SciPhi-Mistral in 81.7% pairs and fine tuned SciPhi-Mistral was preferred to the baseline model in 70.8% pairs. However, responses generated by the fine tuned model were preferred by the curators only slightly less than the ones by GPT-4 model (45.5 versus 54.5%). These results show that the smaller local SciPhi-Mistral-7B-32k can be successfully employed for the description generation task, with the results comparable in quality to the GPT-4 model.

Despite this success, the SciPhi-Mistral-generated annotations are not currently used on the InterPro website. Although SciPhi-Mistral is a cost-effective alternative to GPT-4, the relatively small input context for many families, due to the limited number of Swiss-Prot matches, makes GPT-4's usage manageable in terms of cost.

We plan to continue evaluating both commercial and non-commercial models. As part of this ongoing work, we will always clearly indicate the model used to generate annotations and provide details on the instructions given.

Website and API improvements

Enhancing the integration and visualisation of AlphaFold predicted structures in InterPro

Over 85% of UniProtKB entries now feature predicted protein structures generated by AlphaFold, an artificial intelligence method developed by Google DeepMind that infers protein structures from their amino acid sequences (39,40). In our previous publication, we reported the integration of the Mol* 3D molecular viewer (41) to enable users to visualise predicted structures, which are colour-coded by the predicted local distance difference test (pLDDT), a per-residue confidence measure. The pLDDT score ranges from 0 to 100, with higher scores indicating greater confidence. We have now enhanced our protein viewer to also display the pLDDT confidence scores as a 2D track.

Initially, there were two access points for AlphaFold models. The first is via the 'AlphaFold' tab on the page of a protein with a predicted structure. The second is through the 'AlphaFold' tab on InterPro entry pages, where an example AlphaFold model is shown along with a table listing other available models for that entry. We have now extended the second access point to include all member database signatures, allowing users to access the predicted structures of sequences matched by any member database signature, regardless of whether they have been integrated into an InterPro entry.

Previously, protein structures for Pfam families were predicted using RoseTTAFold (42) and were available for both Pfam families and the InterPro entries in which they were integrated. RoseTTAFold models provided a single, combined view for the family, but if Pfam domain boundaries were incorrect, the structural model would be incomplete. With the majority of Pfam families now having hundreds or thousands of AlphaFold predictions, we have decided to retire the RoseTTAFold models.

Because the sequence associated with a UniProtKB entry may be updated over time, there are instances where an AlphaFold structure was predicted using an earlier version of the sequence. When sequence mismatches are detected, we display a warning message to inform users that the prediction may not accurately represent the current protein structure due to this discrepancy.

With the availability of AlphaFold models for the majority of sequences in UniProtKB, users might be primarily interested in proteins with a predicted structure. Therefore, we have updated all pages listing proteins to indicate whether a protein has a predicted AlphaFold model. When this is the case, we display a link to the page where users can visualise the 3D molecular structure. For example, proteins matched by InterPro can be viewed at https://www.ebi.ac.uk/interpro/protein/UniProt/entry/InterPro/, proteins matched by the Pfam SH2 domain at https://www.ebi.ac.uk/interpro/entry/pfam/PF00017/protein/UniProt/, and

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

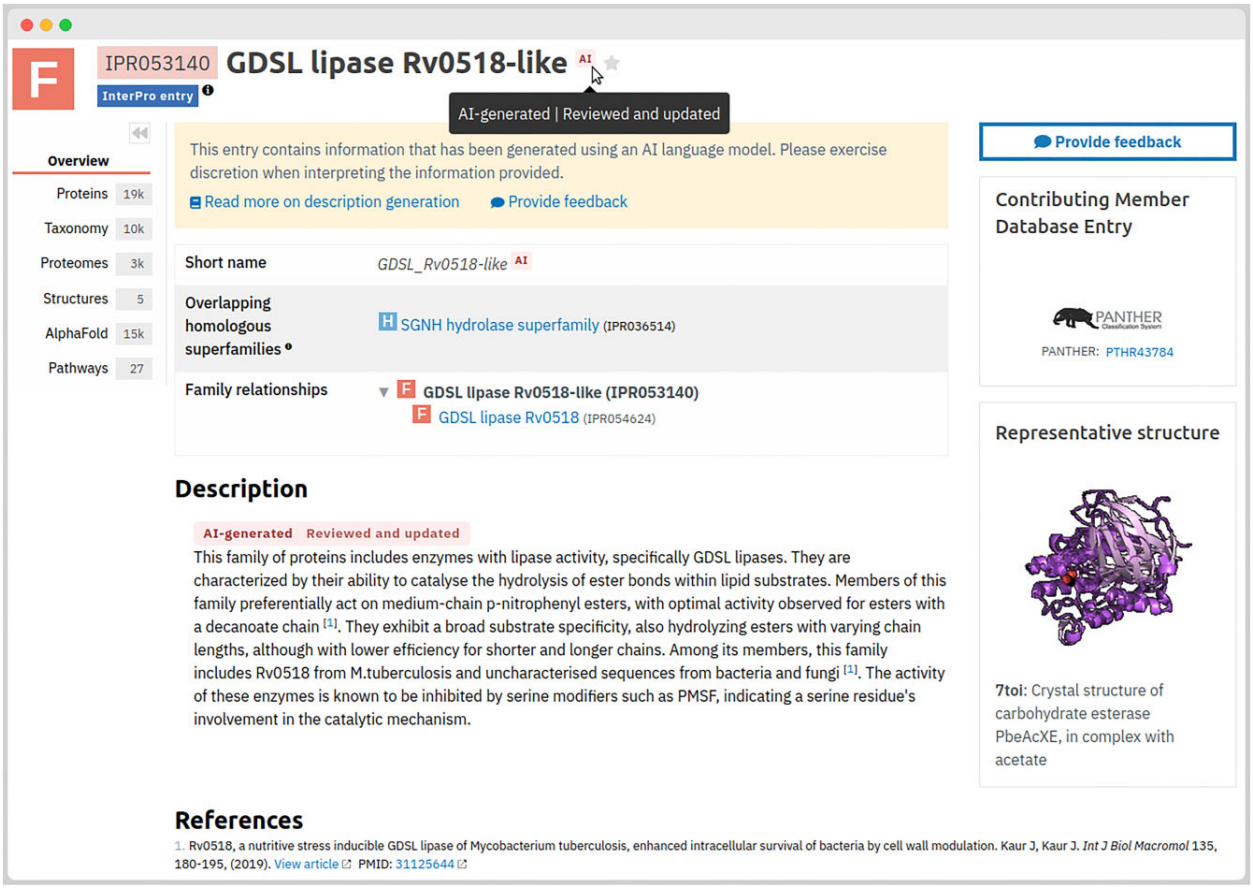


Figure 2. Example of InterPro entry IPR053140, automatically generated using AI annotations for PANTHER family PTHR43784. The entry features an 'AI' label next to the entry name, short name and description to indicate its AI-generated origin. In this instance, a curator has reviewed, updated the description, and included a reference to a relevant scientific publication for supporting evidence. Users are encouraged to use the 'Provide feedback' button to report any inaccuracies or suggest improvements.

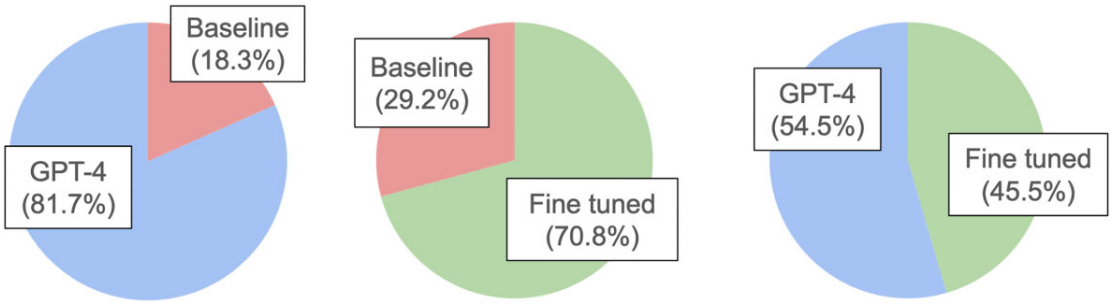


Figure 3. Percentages of pairs where one or another model response was preferred by the InterPro curators.

human proteins at https://www.ebi.ac.uk/interpro/taxonomy/uniprot/9606/protein/UniProt/.

Moreover, enabling structural searches has become a critical feature for researchers. To enhance this capability, we have integrated direct links to the Foldseek webserver (43) on protein pages with AlphaFold predictions and PDB structure pages within InterPro. This one-click access to Foldseek automatically uploads the relevant PDB file to the server making user interactions smoother and faster. By leveraging the extensive structural data now available, this addition significantly improves the use of InterPro for researchers investigating protein function and relationships through structural comparisons.

Improving the visualisation of protein domains and features

Effective visualisation is essential for presenting large volumes of information in a compact, easily interpretable format. InterPro facilitates the visualisation of protein sequence annotations from all its member databases, as well as other regions of interest, such as signal peptides, transmembrane domains, coiled-coils and disordered regions. Although the breadth and diversity of annotations make InterPro a comprehensive resource for protein sequence analysis, they can also lead to potential confusion, especially on protein pages that display over 30 annotation tracks.

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Previously, the Pfam website presented the organisation of domains for a given UniProt entry on a single line. However, Pfam domains do not overlap, and since InterPro integrates domains from multiple member databases, displaying them on a single line would result in overlaps. To enhance clarity and conciseness in visualising protein domains, we now display representative domains at the top of all other tracks in our protein viewer (see Figure 4).

Representative domains are selected through an automated process that compares domains from Pfam, CDD, PROSITE Profiles, SMART and NCBIFAM. The smallest set of domains is chosen to maximise sequence coverage while avoiding overlaps. These representative domains are accessible via the InterPro website, the InterPro REST API and can be downloaded from our FTP in the 'match_complete.xml' file, which includes all matches from member database signatures found in UniProtKB.

Using representative domains selected from several databases offers a more reliable approach to domain annotation. This strategy helps to mitigate and rectify occasional discrepancies that can arise when relying on a single source. For instance, in the case of the human PIK3CA protein, the N345K variant has been reported to lie within the C2 domain (44). However, if one were to rely solely on Pfam domain annotations, the N345K mutation would incorrectly appear outside the C2 domain due to the shorter boundaries of the Pfam C2 domain. By using representative domains, the CDD C2 domain is selected instead, accurately positioning the N345K variant within the C2 domain (see Figure 5). This example highlights the importance of integrating multiple domain databases to ensure accurate variant classification and reduce the risk of misinterpretation.

Since our last publication, we have significantly enhanced the InterPro protein viewer by integrating annotations from several external resources, broadening the scope and depth of the data presented. Among these additions are short linear motifs (SLiMs) sourced from the Eukaryotic Linear Motif (ELM) database (45), a manually curated collection of motif instances derived from experimental literature. SLiMs play crucial roles in mediating protein-protein interactions and other regulatory processes, and their inclusion expands InterPro to include motifs that are often difficult to represent with HMMs.

We have also incorporated data on intrinsically disordered regions from DisProt (46), a comprehensive database that curates experimental evidence of protein disorder from scientific literature. Intrinsically disordered regions are essential for understanding protein flexibility, conformational changes and interactions that do not fit traditional structure-function paradigms.

Additionally, we now display protein repeats retrieved from RepeatsDB (47), a specialised database that annotates tandem repeat structures in proteins. These repeats are often associated with specific functional or structural properties, and their visualisation within the context of InterPro annotations allows for a more detailed analysis of protein architecture.

Disease mutations and variants are integrated into the protein viewer from the Proteins API (48). This integration enables biomedical researchers to correlate protein function and conservation with natural variations and disease-causing mutations. By placing these variants within the broader context of InterPro's extensive protein annotations, researchers can more effectively explore the potential impact of genetic variations on protein function and disease mechanisms.

Lastly, we updated the protein viewer to use the latest version of Nightingale, a library of bioinformatics web components (49). By delegating the visualisation logic to Nightingale, which adheres to strict web standards, we are making the InterPro codebase easier to maintain. Furthermore, the latest version of Nightingale is written in TypeScript, a superset of JavaScript that introduces static types. This enhancement improves code quality and developer experience by providing better tooling for catching errors early and facilitating more robust and readable code.

Enhancing InterProScan's web-based sequence analysis

One of the key features of the InterPro website is the ability for researchers to scan their sequences using InterProScan directly within their web browser. This functionality is particularly beneficial for researchers who may lack access to substantial computing power or the technical expertise required to install and operate command-line tools like InterProScan. It enables users to annotate their sequences seamlessly online. Previously, users were limited to submitting only one protein sequence per search.

The InterProScan search functionality on the InterPro website is powered by the EMBL-EBI Job Dispatcher sequence analysis tools framework. Through collaboration with the team responsible for this framework, we have enhanced the system to allow users to submit up to 100 nucleotide or protein sequences per run. Additionally, we have redesigned the sequence submission, run monitoring, and results pages to streamline the browsing and analysis of results for multiple sequences.

Increasing the integration and accessibility of AntiFam

In our previous publication, we reported the integration of AntiFam into InterProScan, enabling researchers to detect spurious ORFs, a feature particularly useful for quality control in metagenomic and genomic studies. AntiFam hits identified in UniProtKB were displayed alongside other domains and regions of interest within the protein viewer on the InterPro website. Since then, we have enhanced the visibility of AntiFam by allowing users to browse AntiFam families directly and by creating individual pages for each family. This enhancement enables users to explore these families in the same way as other InterPro entries and member database signatures, including the ability to list all sequences in UniProtKB that are matched by a given family. Additionally, we regularly provide UniProt with a list of UniProtKB entries identified as spurious based on AntiFam detection.

InterProScan improvements

Sequence search in InterPro is powered by the InterProScan software. InterProScan takes protein or nucleic acid sequences and searches them against InterPro's predictive models, which are provided by its member databases.

Representative domains

As previously mentioned, the InterPro website now includes representative domains to provide a clear overview of the domain organisation within a sequence. These domains are automatically selected to maximise coverage while minimising

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

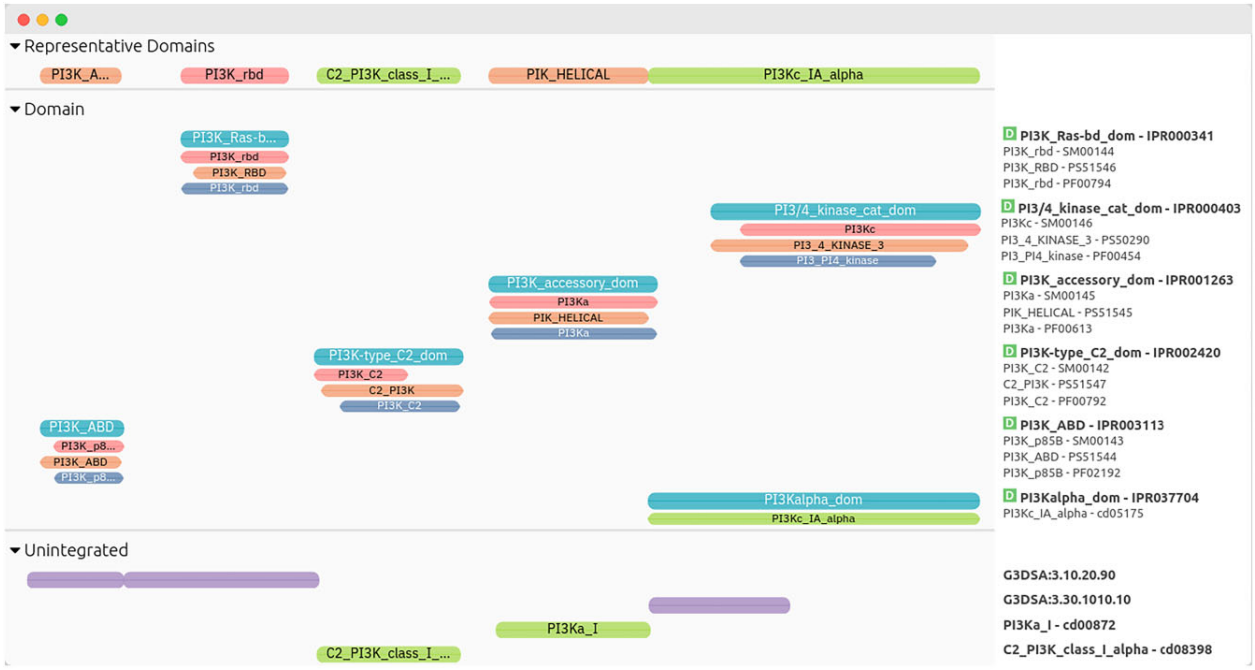


Figure 4. InterPro annotations for the human PIK3CA protein (UniProtKB accession Q15648). The 'Representative Domains' track displays selected domains from InterPro member databases, chosen to maximise coverage and minimise overlap, providing users with a comprehensive overview of the sequence's domain organisation. The 'Domain' section includes InterPro entries classified as domains, along with the underlying member database signatures that match the protein sequence. The 'Unintegrated' section lists annotations from member database signatures that have not yet been integrated into an InterPro entry.

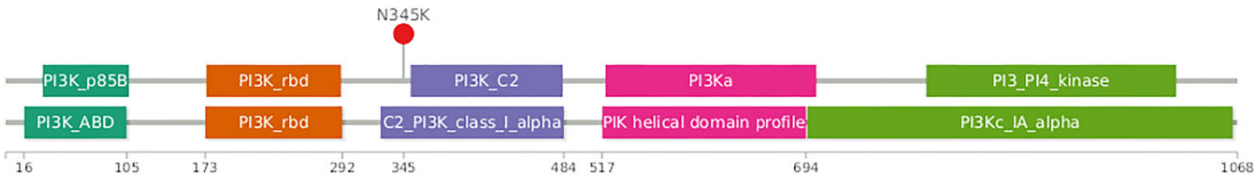


Figure 5. N345K variant in the context of the human PIK3CA protein (UniProtKB accession P42336). The top track illustrates the organisation of Pfam domains, while the bottom track highlights representative domains, with the CDD C2 domain selected over the Pfam C2 domain. Although the N345K mutation, known to reside within the C2 domain, appears outside the Pfam C2 domain, it is located within the boundaries of the CDD C2 domain.

overlaps between domains. InterProScan has been updated to report representative domains in both the JSON and XML output files. A new boolean 'representative' property or attribute has been introduced for each match, allowing users to filter and retain only the representative domains during downstream analyses.

InterProScan docker image for cross-platform compatibility

InterProScan is only supported on Linux systems, but we aim to provide a software that runs out of the box on most personal devices and research institutes' high-performance computing (HPC) clusters. However, many researchers use macOS, which does not natively support InterProScan, and others may have systems that lack the necessary Linux libraries or have outdated versions of these dependencies.

To address these challenges, we created a Docker image of InterProScan, available on DockerHub (https://hub.docker.com/r/interpro/interproscan). Docker allows us to package InterProScan with all required dependencies in a container that can be easily deployed on any system that supports Docker, regardless of the underlying operating system. This ensures that researchers using macOS or systems without the appropriate Linux libraries can still run InterProScan without the need for complex installations or troubleshooting.

Furthermore, for those using containerisation tools like Singularity, SingularityCE, or Apptainer, the Docker image of InterProScan can be pulled directly from DockerHub and used seamlessly, just as it would be with Docker. This ensures that InterProScan is accessible across a wide range of computing environments, including HPC clusters where Singularity is commonly used. By providing a Docker image, we enable broader accessibility and more consistent performance for the research community.

Extending PANTHER annotations

PANTHER has been integrated into InterProScan for over 15 years, traditionally using subfamily HMM scoring to assign sequences to specific subfamilies. However, starting with the release of InterPro 91.0 (October 2022), subfamily HMM scoring was replaced by TreeGrafter (50), a method that annotates protein sequences using pre-annotated phylogenetic

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

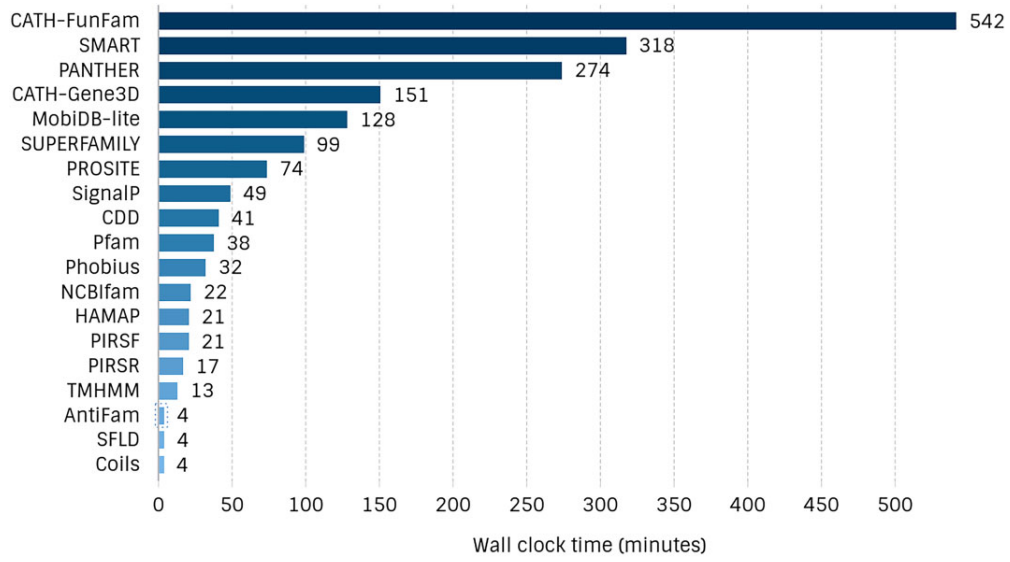


Figure 6. Wall clock time (minutes) consumed by individual member databases to annotate the human proteome using InterProScan with the match lookup disabled. Significant disparities in processing times are evident across the databases, highlighting the computational demands of specific analyses.

trees. TreeGrafter not only outperforms subfamily HMM scoring in accurately assigning subfamily membership, but it also provides highly specific GO term annotations based on these reference phylogenetic trees. InterProScan now reports both the PANTHER family match and the tree graft location, along with subfamily annotations derived from the tree graft.

Starting with InterPro 95.0 (June 2023), GO terms associated with PANTHER tree nodes are reported in InterProScan, similar to how GO terms are associated with InterPro entries (InterPro2GO). These GO annotations are obtained from manually curated, explicit models of function evolution in protein families, from the Phylogenetic Annotation using Gene Ontology (PAN-GO) project (51). This enhances the functional annotation provided by InterProScan, offering researchers more precise annotation of protein function.

Moreover, TreeGrafter is significantly more computationally efficient than subfamily HMM scoring. It takes 13.8 hours to annotate the human proteome with InterProScan using PANTHER 15.0 and subfamily scoring, while it takes only 4.6 h with InterProScan using PANTHER 18.0 and TreeGrafter, demonstrating a useful improvement in efficiency.

Monitoring and improving performances

The performance of InterProScan is largely dependent on the efficiency of the individual binaries used for each member database, the size of these databases (such as the number of profile HMMs), as well as the memory and CPU resources available on the system. When running InterProScan on a HPC system, performance is also influenced by variability in system performance due to factors such as system load, network latency or I/O contention.

Computing is a major contributor to energy consumption, and thus is an important source of carbon emissions. In the context of the global climate crisis, it is imperative that individuals and organisations find ways to assess and then reduce the carbon footprint of their work.

Running InterPro is a computationally intensive activity and, as such, has environmental impacts through carbon dioxide emissions. Between January 2023 and June 2024, we processed over 217 million sequences that were added to UniParc. The resulting annotations are made accessible to InterProScan via a lookup web service, which helps to avoid redundant and wasteful recalculations. When running InterProScan with the match lookup enabled (the default setting), the programme checks whether each input sequence has already been annotated and uses the pre-calculated annotations if available, instead of performing a local scan.

We continuously monitor the compute time of individual member databases within InterProScan to identify opportunities for optimisation and efficiency improvements (see Figure 6). Our goal is to minimise the environmental impact of our computational processes while maintaining the high standards of sequence analysis that our users depend on.

There is a significant variation in the wall clock time consumed by different member databases within InterProScan. MobiDB-lite, a resource focused on predicting intrinsic disorder in proteins, accounted for a compute time of two hours when annotating the human proteome with InterProScan with the match lookup disabled. Although significant, it is not the highest, with CATH-FunFam consuming the most compute time at nine hours. MobiDB-lite generates a consensus by integrating several predictors, such as DisEMBL (52), ESpritz (53) and IUPred (54). MobiDB-lite is implemented in Python 3 and performs system calls to execute the individual predictors, which are distributed as standalone scripts or binaries. This results in a nested and complex series of system calls, coupled with the generation of numerous temporary files, leading to suboptimal performance. Additionally, some predictors require Python 2, which has not been maintained since January 2020, necessitating the installation of both Python 2 and Python 3 on user systems.

To address these inefficiencies, we developed IDRPred (https://doi.org/10.5281/zenodo.13735975), a fast drop-in replacement for MobiDB-lite, and integrated it into Inter-

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

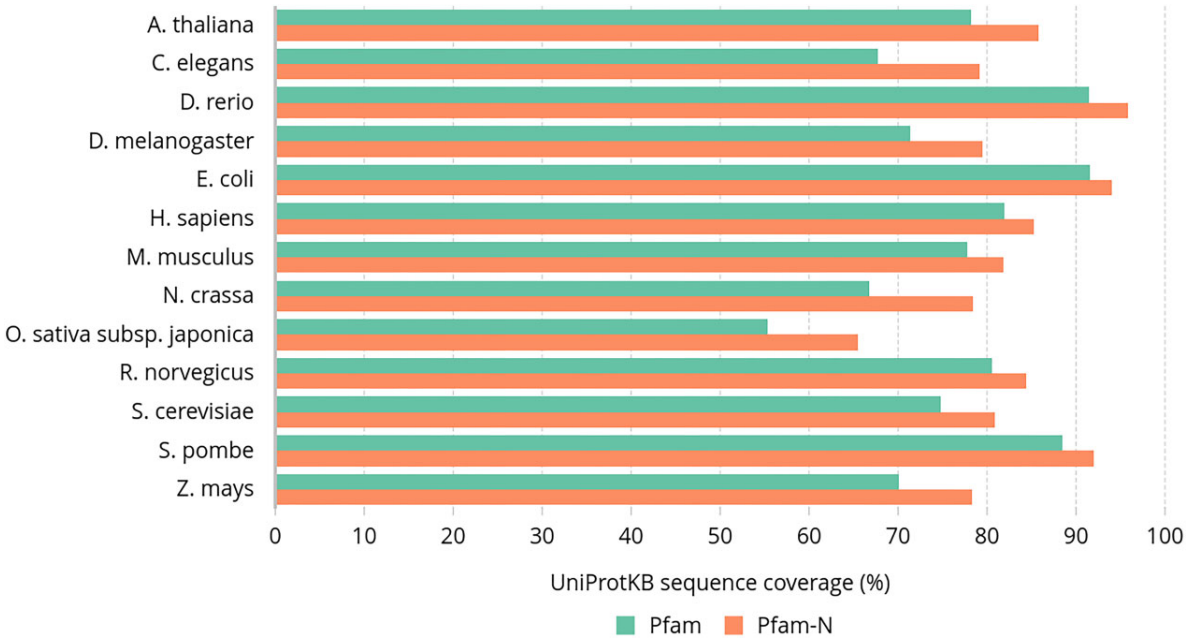


Figure 7. Comparison of annotation coverage between Pfam and Pfam-N across key species and model organisms. For each species, Pfam-N consistently shows higher annotation coverage compared to Pfam, highlighting its enhanced ability to annotate a larger proportion of protein sequences, thereby providing more comprehensive functional information across these organisms.

ProScan. The results are still labelled as MobiDB-lite, and the authors of MobiDB-lite have adopted this fork, which will become the official version. By replacing the original implementation with IDRPred, InterProScan annotates the full human proteome in 90 min instead of 128 min.

By focusing on optimising time-consuming processes, such as those in MobiDB-lite, we can significantly reduce the overall computational footprint of InterProScan, thus contributing to our efforts in lowering carbon emissions associated with large-scale sequence analysis.

Extending protein annotations using deep learning

As part of a collaboration with Google Brain, we integrated ProtENN2 annotations in InterPro 91.0 (released in October 2022). ProtENN2, trained on Pfam data, uses convolutional neural networks to annotate protein sequences with Pfam family labels (*55*). By supplementing the traditional Pfam annotations generated by InterProScan with those provided by ProtENN2, we were able to increase the overall number of sequences that can be labelled with functional annotations. These ProtENN2 annotations, labelled as Pfam-N, were displayed in the InterPro protein viewer, and made accessible via the InterPro REST API and the EMBL-EBI FTP for download.

More recently, Google DeepMind developed a new transformer-based model, which has demonstrated a significant improvement in both prediction accuracy and coverage (Meng-Papaxanthos *et al.*, in preparation). This model was trained on InterPro 100.0 data, including Pfam 37.0. The resulting Pfam-N labels were made available in InterPro with the release of InterPro 101.0 in July 2024.

As with the previous version of Pfam-N, these annotations are now readily accessible through multiple platforms. Users interested in visually exploring protein domains can view the annotations in the InterPro protein viewer. For those who wish to integrate these annotations into their own websites, tools, or downstream analyses, the InterPro REST API provides access. Additionally, users who require annotations for all sequences in UniProtKB can obtain them via bulk download from our FTP site. These multiple access points ensure that the enhanced Pfam-N annotations are available to a broad range of users, from researchers conducting large-scale data analysis to developers building bioinformatics applications.

Pfam-N offers a more comprehensive coverage of UniProtKB sequences compared to Pfam, providing at least one annotation for 85% of sequences in UniProtKB 2024_04, representing an 8% increase over Pfam. It also annotates more sequences than Pfam across several key species and model organisms (see Figure 7). Notably, Pfam-N annotates 12 000 proteins from Swiss-Prot, the reviewed section of UniProtKB, that Pfam does not annotate. Additionally, Pfam-N provides annotations for over 11 million sequences that are not annotated by Pfam or any other InterPro member database, thereby enhancing the comprehensive annotation of known protein sequences.

Conclusion

InterPro continues to evolve as a vital resource for protein sequence classification and functional annotation. The database has made significant progress in integrating new data sources, enhancing analytical capabilities and improving user accessibility.

By bridging the gap between raw sequence data and functional insights, InterPro and InterProScan play a crucial role in advancing our understanding of the protein universe, thereby supporting a wide range of biological research activities.

The field of protein classification has been transformed by the rise of powerful protein structure prediction methods like

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

RosettaFold and AlphaFold. These advances have dramatically improved our ability to predict protein structures, which, in turn, has been instrumental in the creation of Pfam families and families in other databases. These families are subsequently integrated into InterPro entries, enhancing the comprehensiveness and accuracy of the database.

The InterPro REST API remains a popular tool for programmatic access to InterPro data, facilitating its integration into third-party applications and pipelines. However, we have observed an increase in the number of users submitting tens of thousands of requests in a very short period, which can cause the API to become temporarily unavailable for all users. In an era where more public resources are being utilised for AI training, we encourage users to download data in bulk from our FTP site or to start querying our API gradually, then ramp up their usage.

As the volume of sequence data continues to grow rapidly, we have managed to maintain coverage of UniProtKB and UniParc, although we are lagging in expanding this coverage. While achieving 100% sequence or residue coverage is neither a realistic nor sensible goal, we recognise the need to address this challenge. Although the speed of sequence annotation using profile-HMM scoring and other traditional methods is important, it is not currently the main bottleneck in creating InterPro entries and integrating member database signatures. The most significant limiting factor remains the manual curation steps. We have begun using LLMs to annotate member database signatures that lack descriptions, which are then automatically integrated into new InterPro entries. However, this process is dependent on the performance of these LLMs, and the best-performing models are often commercial, limiting our capacity to massively describe protein families and update the generated descriptions.

Ultimately, InterPro relies on the underlying member database signatures it integrates. The coverage of sequence databases by InterPro is, at best, the union of the coverage of its individual member databases. While AI has shown promise in improving our ability to identify homologous in existing protein families, exemplified by Pfam-N and its integration in InterPro, the field is rapidly evolving, and no single model or deep learning architecture has yet emerged as the clear choice for standardisation. Furthermore, while we currently have these annotations available, it is essential that these models be made publicly available and integrated into InterProScan so that end-users can benefit from them.

These are exciting times for the scientific community, and we anticipate significant strides in the coming years towards a more complete and accurate classification of protein families and domains.

Data availability

InterPro releases are published every eight weeks. All data are freely available from the main website (https://www.ebi.ac.uk/interpro), in bulk from the FTP site (https://ftp.ebi.ac.uk/pub/databases/interpro/), and programmatically via the REST API (https://www.ebi.ac.uk/interpro/api/). The latest version of InterProScan can be downloaded from the InterPro website (https://www.ebi.ac.uk/interpro/download/interproscan/). Previous versions are available on the FTP site (https://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/). InterProScan is distributed under the open source Apache 2.0 licence.

Acknowledgements

The work of DH, AMB, NT and FTN was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. We are grateful to Luis Sanchez Pulido, a former member of the InterPro/Pfam team, for his valuable contributions to our curation efforts. We also deeply appreciate the feedback and support from our entire user community.

Funding

Wellcome Trust [221 320/Z/20/Z]; UKRI Biotechnology and Biological Sciences Research Council [BB/S020381/1, BB/X012492/1, BB/X018660/1]; European Molecular Biology Laboratory core funds. European Union - NextGenerationEU through 'Italiadomani-PNRR' project ElixirxNextGenIT [IR0000010] to SCET; National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health [to D.H., A.M.B., N.T. and F.T.N.].

Conflict of interest statement

A.B. is an Editorial Board member of *Nucleic Acids Research*.

References

1. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.* (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
2. Wang,J., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R., Gwadz,M., Lu,S., Marchler,G.H., Song,J.S., Thanki,N., Yamashita,R.A., *et al.* (2023) The conserved domain database in 2023. *Nucleic Acids Res.*, **51**, D384–D388.
3. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S., *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
4. Haft,D.H., Badretdin,A., Coulouris,G., DiCuccio,M., Durkin,A.S., Jovenitti,E., Li,W., Mersha,M., O'Neill,K.R., Virothaisakun,J., *et al.* (2024) RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res.*, **52**, D762–D769.
5. Thomas,P.D., Ebert,D., Muruganujan,A., Mushayahama,T., Albou,L.-P. and Mi,H. (2022) PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.*, **31**, 8–22.
6. Nikolskaya,A.N., Arighi,C.N., Huang,H., Barker,W.C. and Wu,C.H. (2007) PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinform. Online*, **2**, 197–209.
7. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
8. Attwood,T.K., Coletta,A., Muirhead,G., Pavlopoulou,A., Philippou,P.B., Popov,I., Romá-Mateo,C., Theodosiou,A. and Mitchell,A.L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database*, **2012**, bas019.
9. Sigrist,C.J.A., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
10. Akiva,E., Brown,S., Almonacid,D.E., Barber,A.E., Custer,A.F., Hicks,M.A., Huang,C.C., Lauck,F., Mashiyama,S.T., Meng,E.C.,

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

et al. (2013) The structure–function linkage database. *Nucleic Acids Res.*, **42**, D521–D530.
11. Letunic,I., Khedkar,S. and Bork,P. (2021) SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.*, **49**, D458–D460.
12. Oates,M.E., Stahlhacke,J., Vavoulis,D.V., Smithers,B., Rackham,O.J.L., Sardar,A.J., Zaucha,J., Thurlby,N., Fang,H. and Gough,J. (2014) The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res.*, **43**, D227–D233.
13. Ontology Consortium,G., Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., *et al.* (2023) The gene ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
14. Burge,S., Kelly,E., Lonsdale,D., Mutowo-Muellenet,P., McAnulla,C., Mitchell,A., Sangrador-Vegas,A., Yong,S.-Y., Mulder,N. and Hunter,S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database*, **2012**, bar068.
15. Necci,M., Piovesan,D., Clementel,D., Dosztányi,Z. and Tosatto,S.C.E. (2021) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, **36**, 5533–5534.
16. Nielsen,H. (2017) Predicting secretory proteins with SignalP. *Methods Mol. Biol.*, **1611**, 59–73.
17. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.
18. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
19. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
20. Das,S., Lee,D., Sillitoe,I., Dawson,N.L., Lees,J.G. and Orengo,C.A. (2015) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics*, **31**, 3460–3467.
21. Eberhardt,R.Y., Haft,D.H., Punta,M., Martin,M., O'Donovan,C. and Bateman,A. (2012) AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database*, **2012**, bas003.
22. Drysdale,R., McEntyre,J., Durinx,C. and Blomberg,N. (2018) The process for the selection of ELIXIR core data resources. *F1000Res.*, **7**, 1712.
23. Anderson,W.P. and Global Life Science Data Resources Working GroupGlobal Life Science Data Resources Working Group (2017) A global coalition to sustain core data. *Nature*, **543**, 179.
24. Madeira,F., Madhusoodanan,N., Lee,J., Eusebi,A., Niewielska,A., Tivey,A.R.N., Lopez,R. and Butcher,S. (2024) The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.*, **52**, W521–W525.
25. Consortium,U.P. (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
26. MacDougall,A., Volynkin,V., Saidi,R., Poggioli,D., Zellner,H., Hatton-Ellis,E., Joshi,V., O'Donovan,C., Orchard,S., Auchincloss,A.H., *et al.* (2021) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics*, **36**, 5562.
27. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
28. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O., Azov,A.G., Barba,M., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2024) Ensembl 2024. *Nucleic Acids Res.*, **52**, D891–D899.
29. Yates,A.D., Allen,J., Amode,R.M., Azov,A.G., Barba,M., Becerra,A., Bhai,J., Campbell,L.I., Martinez,C., M.,C., *et al.*2022) Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.*, **50**, D996–D1003.
30. Armstrong,D.R., Berrisford,J.M., Conroy,M.J., Gutmanas,A., Anyango,S., Choudhary,P., Clark,A.R., Dana,J.M., Deshpande,M., Dunlop,R., *et al.* (2020) PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.*, **48**, D335–D343.
31. Conesa,A. and Götz,S. (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.
32. Richardson,L.J., Rawlings,N.D., Salazar,G.A., Almeida,A., Haft,D.R., Ducq,G., Sutton,G.G. and Finn,R.D. (2019) Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. *Nucleic Acids Res.*, **47**, D564–D572.
33. Richardson,L., Allen,B., Baldi,G., Beracochea,M., Bileschi,M.L., Burdett,T., Burgin,J., Caballero-Pérez,J., Cochrane,G., Colwell,L.J., *et al.* (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, **51**, D753–D759.
34. Gurbich,T.A., Almeida,A., Beracochea,M., Burdett,T., Burgin,J., Cochrane,G., Raj,S., Richardson,L., Rogers,A.B., Sakharova,E., *et al.* (2023) MGnify genomes: a resource for biome-specific microbial genome catalogues. *J. Mol. Biol.*, **435**, 168016.
35. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
36. Green,A., Ribas,C., Ontiveros-Palacios,N., Griffiths-Jones,S., Petrov,A.I., Bateman,A. and Sweeney,B. (2023) LitSumm: large language models for literature summarisation of non-coding RNAs. arXiv doi: https://arxiv.org/abs/2311.03056, 06 November 2023, preprint: not peer reviewed.
37. Open,A.I., Achiam,J, Adler,S., Agarwal,S., Ahmad,L., Akkaya,I., Aleman,F.L., Almeida,D., Altenschmidt,J., Altman,S., *et al.* (2023) GPT-4 technical report. arXiv doi: https://arxiv.org/abs/2303.08774, 15 March 2023,preprint: not peer reviewed.
38. Jiang,A.Q., Sablayrolles,A., Mensch,A., Bamford,C., Chaplot,D.S., de las Casas,D., Bressand,F., Lengyel,G., Lample,G., Saulnier,L., *et al.* (2023) Mistral 7B. arXiv doi: https://arxiv.org/abs/2310.06825, 10 October 2023, preprint: not peer reviewed.
39. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
40. Varadi,M., Bertoni,D., Magana,P., Paramval,U., Pidruchna,I., Radhakrishnan,M., Tsenkov,M., Nair,S., Mirdita,M., Yeo,J., *et al.* (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.*, **52**, D368–D375.
41. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
42. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Dustin Schaeffer,R., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
43. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2023) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
44. Gymnopoulos,M., Elsliger,M.-A. and Vogt,P.K. (2007) Rare cancer-specific mutations in PIK3CA show gain of function. *Proc. Natl. Acad. Sci.*, **104**, 5569–5574.
45. Kumar,M., Michael,S., Alvarado-Valverde,J., Mészáros,B., Sámano-Sánchez,H., Zeke,A., Dobson,L., Lazar,T., Örd,M., Nagpal,A., *et al.* (2021) The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.*, **50**, D497–D508.

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Högskolan Biblioteket user on 02 October 2026

46. Aspromonte,M.C., Nugnes,M.V., Quaglia,F., Bouharoua,A., Tosatto,S.C.E. and Piovesan,D. (2023) DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res.*, **52**, D434–D441.
47. Paladin,L., Bevilacqua,M., Errigo,S., Piovesan,D., Mičetić,I., Necci,M., Monzon,A.M., Fabre,M.L., Lopez,J.L., Nilsson,J.F., *et al.* (2020) RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res.*, **49**, D452–D457.
48. Nightingale,A., Antunes,R., Alpi,E., Bursteinas,B., Gonzales,L., Liu,W., Luo,J., Qi,G., Turner,E. and Martin,M. (2017) The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res.*, **45**, W539–W544.
49. Salazar,G.A., Luciani,A., Watkins,X., Kandasaamy,S., Rice,D.L., Blum,M., Bateman,A. and Martin,M. (2023) Nightingale: web components for protein feature visualization. *Bioinform. Adv.*, **3**, vbad064.
50. Tang,H., Finn,R.D. and Thomas,P.D. (2018) TreeGrafter: phylogenetic tree-based annotation of proteins with gene ontology terms and other annotations. *Bioinformatics*, **35**, 518–520.
51. Gaudet,P., Livstone,M.S., Lewis,S.E. and Thomas,P.D. (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief Bioinform*, **12**, 449–462.
52. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
53. Walsh,I., Martin,A.J.M., Di Domenico,T. and Tosatto,S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
54. Mészáros,B., Erdős,G. and Dosztányi,Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, **46**, W329–W337.
55. Bileschi,M.L., Belanger,D., Bryant,D.H., Sanderson,T., Carter,B., Sculley,D., Bateman,A., DePristo,M.A. and Colwell,L.J. (2022) Using deep learning to annotate the protein universe. *Nat. Biotechnol.*, **40**, 932–937.

Downloaded from academic.oup.com/nar/article/53/D1/D444/7905301 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Received: September 13, 2024. **Revised:** October 11, 2024. **Editorial Decision:** October 14, 2024. **Accepted:** October 23, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

D520–D528 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 24 October 2018*
doi: 10.1093/nar/gky949

Protein Data Bank: the single global archive for 3D macromolecular structure data

wwPDB consortium

Received September 14, 2018; Revised September 28, 2018; Editorial Decision October 01, 2018; Accepted October 05, 2018

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

ABSTRACT

The Protein Data Bank (PDB) is the single global archive of experimentally determined three-dimensional (3D) structure data of biological macromolecules. Since 2003, the PDB has been managed by the Worldwide Protein Data Bank (wwPDB; wwpdb.org), an international consortium that collaboratively oversees deposition, validation, biocuration, and open access dissemination of 3D macromolecular structure data. The PDB Core Archive houses 3D atomic coordinates of more than 144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules and related experimental data and metadata. Structure and experimental data/metadata are also stored in the PDB Core Archive using the readily extensible wwPDB PDBx/mmCIF master data format, which will continue to evolve as data/metadata from new experimental techniques and structure determination methods are incorporated by the wwPDB. Impacts of the recently developed universal wwPDB OneDep deposition/validation/biocuration system and various methods-specific wwPDB Validation Task Forces on improving the quality of structures and data housed in the PDB Core Archive are described together with current challenges and future plans.

INTRODUCTION

The Protein Data Bank (PDB, pdb.org) was established in 1971 as the first open-access, molecular data resource in biology (1). More than 47 years later, the PDB continues to serve as the single global repository for atomic-level, 3D structure data, making >144 000 experimentally-determined structures of proteins, DNA, and RNA, and their complexes with metal ions, drugs, and other small molecules freely available without restrictions on use. Since 2003, the PDB has been managed jointly by the Worldwide Protein Data Bank (wwPDB) consortium (2), including the US Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB; rcsb.org) (3), the Protein Data Bank in Europe (PDBe; pdbe.org) (4), Protein Data Bank Japan (PDBj; pdbj.org) (5) and BioMagResBank (BMRB; www.bmrb.wisc.edu) (6). The wwPDB partners are committed to ensuring adherence to the **FAIR** Principles of **F**indability-**A**ccessibility-**I**nteroperability-**R**eusability (7).

Today, the PDB is universally regarded as a core data resource essential for understanding the functional roles that macromolecules play in biology and medicine. Publication of new macromolecular structures in most scientific journals is contingent on mandatory deposition to the PDB of the 3D atomic coordinates comprising the structural model plus experimental data used to derive the structures and associated metadata. Many governmental and non-governmental research funders also require PDB deposition of unpublished macromolecular structure data. All of these 3D structural data are stored in one of two wwPDB Core Archives. The PDB Core Archive houses 3D atomic coordinates of >144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules. The PDB Core Archive also houses related experimental data/metadata from Macromolecular Crystallography (MX). The BioMagResBank (BMRB; www.bmrb.wisc.edu) Core Archive houses related experimental data/metadata from Nuclear Magnetic Resonance spectroscopy (NMR). The wwPDB partners work closely with the Electron Microscopy Data Bank (EMDB; emdb-empiar.org), which houses related experimental data/metadata from 3D Electron Microscopy (3DEM) and Electron Tomography (ET).

The PDB Core Archive has seen steady growth since its inception, with over 11,000 new structures plus experimental data/metadata released in 2017 (Figure 1A). In aggregate, most of the 3D structures (89.5%) in the PDB Core Archive were determined using macromolecular crystallography (MX), with the remainder determined by NMR (8.5%), 3DEM (1.6%), and other techniques (0.4%). These overall metrics mask recent trends, which show that in 2016 3DEM overtook NMR as the second most popular technique for determining atomic level structures (Figure 1B).

While the PDB Core Archive has grown enormously in scale and scope over the past 47 years and its management has evolved concurrently, adherence to the principle of open access and commitment to community engagement (1) continue to this day.

The Vision of the wwPDB is to:

- Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macro-

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

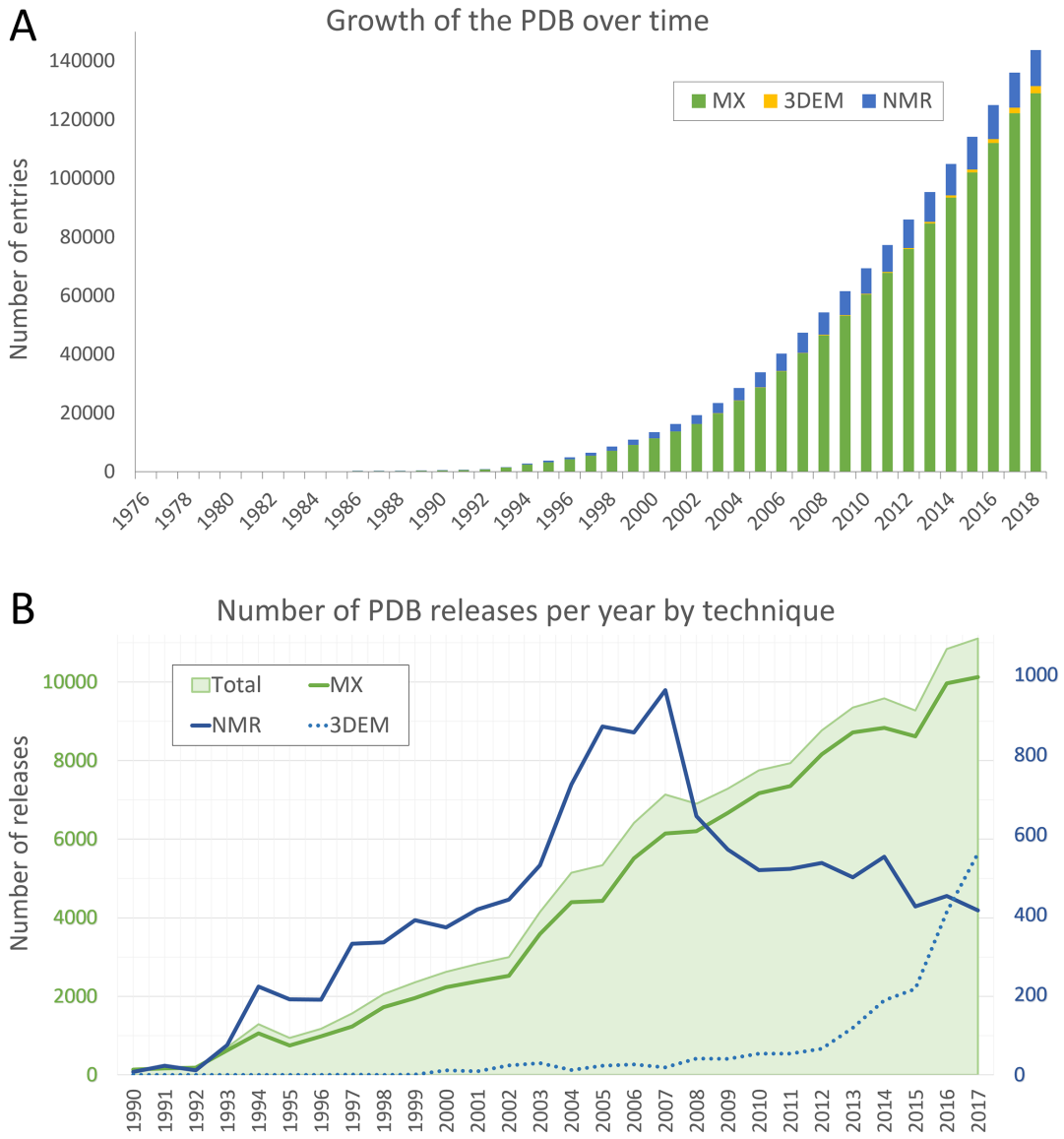


Figure 1. (**A**) Growth the PDB Core Archive. Total height of each bar indicates aggregate released structures, coloured by experimental technique (MX—green, 3DEM—yellow, NMR—blue). (**B**) Number of PDB structures released annually. All PDB Core Archive structures are indicated with light green shading, and MX structures are shown with a solid green line, plotted with respect to the green primary axis (left). NMR structures (blue solid line) and 3DEM structures (blue dashed line) are plotted with respect to the blue secondary axis (right).

molecules as an enduring public good to promote basic and applied research and education across the sciences.

The Mission of the wwPDB is to:

- Manage the wwPDB Core Archives as a public good according to the **FAIR** Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors worldwide.
- Ensure universal open access to public domain structural biology data with no limitations on usage.
- Develop and promote community-endorsed data standards for archiving and exchange of global structural biology data.

PDB data are being used by researchers, educators, specialist bioinformatics resources and other Users from every inhabited continent and every UN-recognized sovereign nation. Nearly two million daily structure data file downloads from wwPDB partner websites and the two Core Archives attest to the important role that the wwPDB plays within the biological data ecosystem.

Ongoing collaborative work among wwPDB partners helps to ensure completeness, consistency, and accuracy of data in the two Core Archives. The wwPDB has also worked to enable growth of the corpus of structure data to accommodate new experimental techniques. Herein, we describe impacts of the recently developed universal wwPDB OneDep deposition/validation/biocuration sys-

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

tem and various methods-specific wwPDB Validation Task Forces on improving the quality of structures and data housed in the PDB Core Archive, together with current challenges and future plans.

PDB CORE ARCHIVE CONTENT

Unlike the situation in 1971, multiple techniques are now available for determining 3D structures of biological macromolecules. PDB structure depositions are currently restricted to atomic-level structures that have been substantially determined by one or more of the following supported experimental techniques: MX, NMR, 3DEM, powder diffraction and fiber diffraction.

Atomic coordinate data

Every PDB structure deposition includes the atomic coordinates defining the 3D structural model of the macromolecule. Atomic positions are specified as Cartesian coordinates (x, y, z) using Ångström units (i.e. 0.1 nm) and a right-handed coordinate system. Additional method-specific attributes are provided for individual atoms (e.g. *B*-factors or temperature-factors for MX structures).

Related metadata

To ensure adherence to the **FAIR** Principles (7), the atomic coordinates of 3D structures must be adequately described with additional mandatory metadata. These metadata include a hierarchy of information describing whether a particular atom is part of a polymer (and if so, which residue), or a metal ion, a ligand, a small molecule solute or a water molecule. For macromolecules, additional metadata including name, source organisms, and cross-references to other bioinformatics resources are provided. Data on the type of structure determination experiment performed and on the nature and production of the experimental sample are also archived. Consistent collection of structure data and experimental data/metadata, governed by defined vocabularies, allows Users of the PDB Core Archive to find and understand 3D structures of interest.

Experimental data

Deposition of experimental data/metadata together with atomic coordinate data is required for all incoming structures. For MX experiments, deposition of structure factors or unmerged intensities and related metadata is required. These data are stored in the PDB Core Archive. For NMR experiments, deposition of assigned chemical shifts and geometric restraints and related metadata is required. These data and additional experimental data are stored in the BMRB Core Archive. For 3DEM experiments, Coulomb potential maps and related metadata are required. These data are stored in the EMDB (8). Experimental data/metadata accompanying MX, NMR and 3DEM structures are processed using the universal OneDep system for deposition/validation/biocuration (9). For NMR and 3DEM methods wherein experimental data are often deposited in advance of 3D structure determination, cross-referencing with subsequently deposited atomic coordinates ensures interoperability across the PDB and BMRB Core Archives and the EMDB. The OneDep system also allows Data Depositors to provide additional links to other experimental data, housed in repositories such as SBGRID (sbgrid.org) (10), IIRMC (proteindiffraction.org) (11), SASBDB (www.sasbdb.org) (12) and EMPIAR (www.ebi.ac.uk/pdbe/emdb/empiar) (13).

Mandatory archiving of the experimental data fulfils two important functions. First, it enables PDB Users to reproduce 3D structure determinations and analyses therefrom. Second, it allows quantitative assessments of how well the atomic coordinates conform to the experimental data. The wwPDB partnership has made significant investment in validation of 3D atomic-level structures together with experimental data/metadata (9, 14–19).

Chemical reference data

In addition to atomic coordinates and experimental data, the wwPDB provides key chemical reference data including the Chemical Component Dictionary (CCD) (20) and Biologically Interesting Molecule Reference Dictionary (BIRD) (21). The CCD provides a detailed chemical description of every unique chemical component represented within 3D atomic coordinates in the PDB Core Archive, including standard and modified residues, metal ions, small molecule ligands, solute molecules, and water molecules. Each chemical component definition includes descriptions of chemical properties, such as stereochemical assignments, chemical structure descriptors (SMILES and InChI), systematic chemical names, chemical formulae, and idealized atomic coordinates. Currently, the BIRD includes detailed descriptions of biologically interesting peptide-like antibiotic and enzyme inhibitor molecules present in the PDB Core Archive. These molecules may be composed of a mixture of polymer and non-polymer components or short polymeric entities, and require a description on the level of the whole molecule and on the level of constituent parts. In future, the BIRD resource could be extended to other kinds of oligomeric molecules, which may require analogous dual definitions.

PDBx/mmCIF DATA FILE FORMAT

Providing consistent and accurate representation for all 3D structures in the PDB Core Archive allows these data to be easily searched and exploited by Users around the world. Significant advances in structure determination techniques over the last decade have resulted in an increase in the size and complexity of macromolecular structures studied by structural biologists. As a result, it is no longer possible to represent these large macromolecular machines with the legacy PDB file format, which is restricted to a maximum of 99 999 atoms and 62 single-character polymer chain identifiers. For a time, 3D structural models that exceeded these limits were split into multiple PDB entries, causing considerable inconvenience to Data Depositors and Users alike. To address this limitation, the wwPDB convened a working group to obtain community support for adoption of a common extensible PDBx/mmCIF data archiving framework (22–24) with the associated mmCIF format as the master file format for the PDB Core Archive. Large structures

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

that were previously split across multiple PDB entries were merged into a single PDB entry using the PDBx/mmCIF format. All 3D structures in the PDB Core Archive are now stored and distributed in PDBx/mmCIF format.

Where possible, the wwPDB has continued to make structures in the PDB Core Archive available in legacy PDB format for the convenience of the User community. For the avoidance of doubt, the legacy PDB format was 'frozen' in 2012, and no longer conveys the broad range of rich metadata represented in the PDBx/mmCIF files. Most major structural biology data resources and software tools have embraced the PDBx/mmCIF format, and continued reliance on the legacy PDB file format is strongly discouraged by the wwPDB.

Experimental data are distributed using various file formats, reflecting minor differences in the evolution of data archiving practices within different structural biology communities. Structure factors from MX experiments are stored and distributed in the PDBx/mmCIF format. Chemical shifts from NMR experiments are stored and distributed in the BMRB Core Archive NMR-STAR format (25, 26). Coulomb potential maps from 3DEM experiments are stored and distributed in the CCP4 map format, whereas the associated metadata are stored and distributed in EMDB-XML format (27).

All the data categories and items represented using the PDBx/mmCIF format are unambiguously defined in the PDBx/mmCIF dictionary (mmcif.wwpdb.org) (22). These data definitions include relationships among different data categories, and allowed data types for all data items in the data categories. The PDBx/mmCIF dictionary also includes allowed enumerations and ranges of values for individual data items, thereby enabling validation of the data items in each PDBx/mmCIF file across the PDB Core Archive. Finally, the PDBx/mmCIF format has the advantage of being fully extensible, enabling archiving of new data types as structural biology continues to develop as a scientific discipline.

Additional structure data file formats

In addition to PDBx/mmCIF, the wwPDB also distributes the atomic coordinates of every PDB structure in PDBML/XML format, which can be read using a standard XML parser (pdbml.pdb.org) (28) and as RDF (rdf.wwpdb.org/pdb) (5,29).

GLOBAL DATA DEPOSITION

In 2014, the global OneDep deposition-validation-biocuration system was launched (9). This important advance ensured that all wwPDB regional data centers (RCSB PDB, PDBe and PDBj) provide the same data processing experience to Data Depositors around the world. The OneDep system processes depositions of 3D structures coming from all of the experimental methods currently supported by PDB Core Archives. OneDep also supports all EMDB depositions, including Coulomb potential maps that are not accompanied by 3D atomic coordinates. Validation and biocuration of depositions is geographically distributed, with RCSB PDB processing all depositions

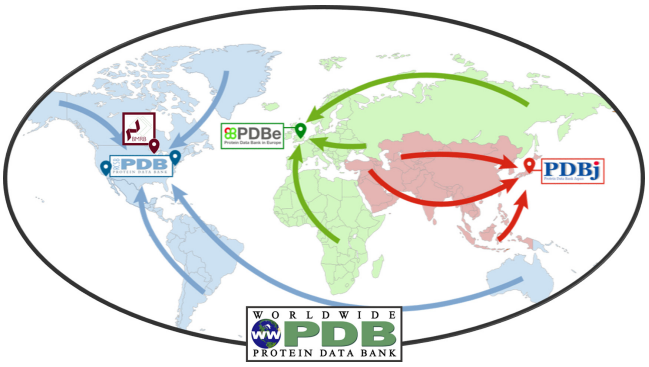


Figure 2. World map indicating the locations of wwPDB partner sites and color-coded to indicate the regions from which each accepts PDB depositions.

from the Americas and Oceania, PDBe processing depositions from Europe and Africa, and PDBj processing all depositions from Asia and the Middle East (Figure 2). This arrangement distributes validation/biocuration efforts across three wwPDB partner sites, and allows most Data Depositors to communicate with wwPDB biocurators located in the same or nearby time zones.

To improve data quality and to allow for future extensibility, the OneDep system uses the PDBx/mmCIF framework throughout deposition-validation-biocuration. The launch of the OneDep system was accompanied by a significant extension of the PDBx/mmCIF dictionary, including addition of new or updated enumerations and allowed ranges for individual data items and significant changes to improve representation of 3DEM structures within the PDB Core Archive. Both the PDBx/mmCIF dictionary and the OneDep system undergo continuous updates, with enumerations being extended and new data items being added when and where appropriate.

wwPDB validation (15) and biocuration (30) processes have been described in detail, and information about the deposition and biocuration policies and procedures is available on the wwPDB website (www.wwpdb.org/documentation/biocuration). Recently, validation of ligands in the PDB Core Archive was improved with the adoption of a more robust way of flagging those molecules that do not fit electron density well (31). The wwPDB validation report provides consistent quality assessment metrics across the entire PDB archive, enabling comparisons between different structures. This allows Users to rank PDB structures relevant for their needs based on validation criteria.

Validation is performed throughout the deposition process: a preliminary validation report is provided to Data Depositors at the time of deposition; an updated, confidential report is then provided to Data Depositors once the biocuration is concluded; and a final public report is released alongside the PDB entry. An increasing number of scientific journals now require wwPDB validation reports to be included at the time of manuscript submission to assist referees in assessment of the quality of the 3D structure data. The wwPDB strongly encourages all Data Depositors to provide their confidential wwPDB validation reports when submitting related manuscripts.

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Release of PDB entries requires either a request from the Data Depositor, a notification that the related manuscript has been published in a scientific journal, or the expiration of the on-hold period (currently maximum one year from the date of deposition), whichever occurs first.

The RCSB PDB currently serves as the designated PDB Core Archive Keeper, coordinating a two-phase weekly update procedure for both new and revised entries. Each wwPDB regional data center finalizes entries for the next weekly release up until Thursday at noon local time. At this time, data marked for release are compiled and re-checked at RCSB PDB and then distributed to all wwPDB partner sites. Phase 1 of the weekly update occurs every Saturday at 03:00 UTC, when the wwPDB website (www.wwpdb.org/download/downloads) makes the following data available for each new entry: amino acid and nucleotide sequences for each distinct polymer and, where appropriate, the InChI string(s) for each distinct ligand, and the crystallization pH value(s). Phase 2 of the weekly update occurs every Wednesday at 00:00 UTC, when the wwPDB FTP system and the FTP sites at wwPDB partner sites are updated to include all new and modified PDB entries and obsolete entries are removed from the active archive. Release of PDB data in two phases is intended to assist the computational biology community in operating various prediction challenges, including CASP (32), CAPRI (33), CAMEO (34) and D3R (35).

DATA DISSEMINATION

The wwPDB website (www.wwpdb.org) provides news announcements, describes how to access PDB data, and hosts PDBx/mmCIF data dictionary resources.

PDB data are made available from all wwPDB partner sites *via* FTP and also through their individual websites - RCSB PDB (rcsb.org), PDBe (pdbe.org), PDBj (pdbj.org), and BMRB (www.bmrb.wisc.edu). In 2017, the FTP archive recorded >450 million structure data file downloads, and at the individual wwPDB partner websites they numbered >220 million (Figure 3). The aggregate number of unique PDB Users (unique IP addresses) worldwide is conservatively estimated at >1 million.

Following consultation with PDB Users, the wwPDB implemented versioning of data in the PDB Core Archive. Automatic auditing of changes to PDB entries has been introduced, which distinguishes between updates to atomic coordinates, chemistry or polymer sequence (denoted as 'major') and other updates, including citation updates (denoted as 'minor'). A versioned FTP archive (ftp://ftp-versioned.wwpdb.org) has been introduced, which serves up the latest minor version of each major version of a PDB entry. This provision allows Users access to information on updates to each PDB structure and allows for comparison of available major versions to review changes. In the versioned FTP, PDB entries are identified by an eight character ID allowing for extension of the PDB code beyond its current four characters.

All MX experimental data in the PDB Core Archive are distributed *via* the wwPDB FTP. Experimental data relating to 3D structures coming from NMR are also distributed *via* the wwPDB FTP. Additional NMR experimental data associated with 3D structures, not collected by OneDep but deposited at BMRB, are distributed by BMRB. The wwPDB FTP also mirrors the EMDB FTP, providing access to the entire contents of the EMDB.

ARCHIVE UPDATES

To improve consistency of data within the PDB Core archive, wwPDB biocurators routinely update PDB entries with new or corrected metadata, such as citation information or updates triggered by changes to the CCD (30).

The wwPDB has undertaken several archive-wide remediations, starting with the standardization of the PDB Core archive in 2007. This major undertaking introduced chemical descriptors to the CCD and ensured that atom names of standard residues are consistent with IUPAC nomenclature (36). Subsequent remediations have focused on adding further metadata, including taxonomy information (2009), and on introducing missing molecular assembly information (2011). As noted above, in 2014, after the adoption of PDBx/mmCIF as the master format, structural models that spanned more than one PDB entry due to limitations of the historical PDB file format were combined into single PDB entries distributed exclusively in PDBx/mmCIF format.

Launch of the OneDep system has enabled, for the first time, large-scale remediation to improve data consistency across all entries in the PDB Core archive. Older PDB entries deposited using legacy deposition systems were updated to ensure that their metadata are consistent with newer entries deposited using the OneDep system (~30% of older PDB entries underwent remediation). This effort also included better representation of data related to 3DEM entries. Remediation is an ongoing process, intended to ensure better data quality and consistency and improved searchability across the PDB Core Archives. Consistent representation of carbohydrates and post-translational modifications will be the focus of future remediation efforts.

COMMUNITY ENGAGEMENT

The enduring value of the PDB Core archive to its large, global User community depends critically on data consistency, accuracy, and accessibility. It is, therefore, essential to ensure that the User needs are both understood and addressed. The wwPDB is guided by an expert international advisory board, which reviews the activities of the organization annually to ensure that the wwPDB is delivering value to its diverse User community. Outcomes of the annual advisory board meetings are published on the wwPDB website (www.wwpdb.org/about/advisory). This advisory board provides essential guidance for wwPDB developments and delivers community feedback on changes that can benefit the PDB archive.

The field of structural biology is constantly evolving, and the wwPDB is committed to staying abreast of these advances. Over the years, expert, method-specific wwPDB Validation Task Forces have been established for MX (19), NMR (17) and 3DEM (16). Each of these groups contributed to the development of the OneDep validation system and the wwPDB validation report. In collaboration with the Cambridge Crystallographic Data Centre

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

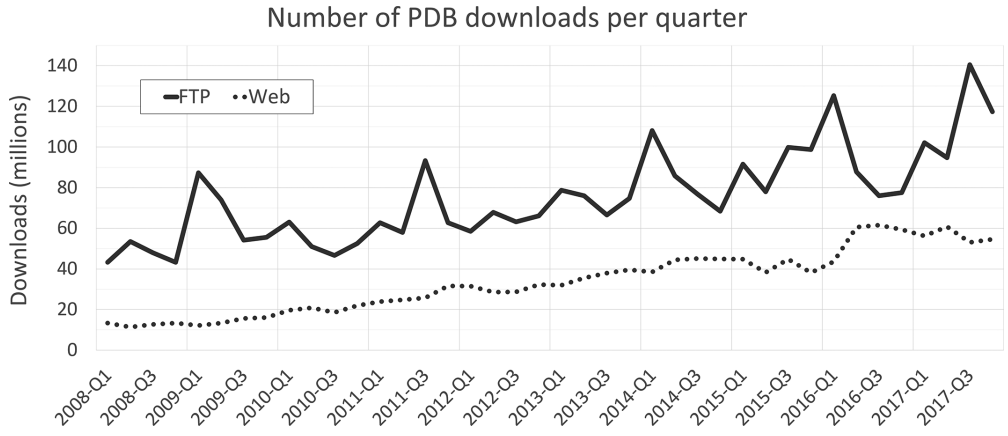


Figure 3. Quarterly wwPDB structure data file download metrics from 2008–2017 (FTP and rsync downloads-solid line; wwPDB partner website downloads-dotted line).

(37) and the Drug Design Data Resource (D3R) (www.drugdesigndata.org), the wwPDB convened a Ligand Validation Workshop in 2015 to obtain recommendations regarding improved PDB ligand representation and validation (18). Further refinements of our validation processes are being guided by continued interaction with the validation task forces. The wwPDB validation report is updated annually to incorporate software updates, new validation processes, and to update the archive wide validation statistics by incorporating PDB entries from the previous year.

To establish data standards and obtain recommendations for improving data quality on rapidly developing experimental methods currently not supported within the PDB archive, the wwPDB has also established task forces for Small Angle Scattering (SAS) (38) and Integrative/Hybrid Methods (I/HM) (39), both of which have published their recommendations in white papers.

All of these efforts are underpinned by the PDBx/mmCIF Working Group (www.wwpdb.org/task/mmcif), which advises on data standards for representation of structural biology data in the PDBx/mmCIF dictionary. This working group meets regularly with representatives from the wwPDB to advise on adjustments to the data dictionary.

In consultation with the working group, the wwPDB has recently produced an extension to the PDBx/mmCIF dictionary to incorporate multiple crystal data collection techniques such as those used in serial femtosecond crystallography (SFX) and X-ray free electron laser (XFEL) experiments. This step increased the amount of metadata that will be available for future multiple crystal data collection experiments. None of this would have been possible had the organization not transitioned from the legacy PDB format to the PDBx/mmCIF data standard. The wwPDB will continue to expand and extend the PDBx/mmCIF dictionary as structural biology advances.

The wwPDB has also worked with the NMR community to develop the NMR Exchange Format (NEF) (40), and this format for deposition of NMR restraint data will be supported by the OneDep system in a later software release.

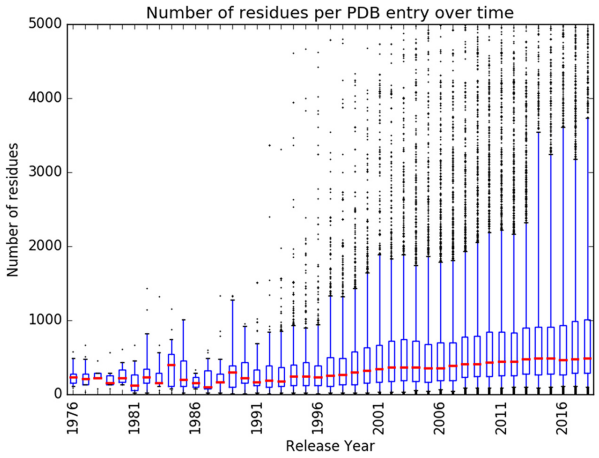


Figure 4. Boxplot representation of the number of residues per PDB structure per year. Red lines represent the median value, boxes represent the 25 and 75 percentile values, and whiskers represent the 5th and 95th percentile values.

Communications from Data Depositors regarding their own depositions should only be submitted via the Communication Panel within the individual OneDep session. More general feedback regarding issues not directly related to an individual OneDep session is welcome at the wwPDB Customer Service Helpdesk—further contact information is available at deposit.wwpdb.org.

CURRENT AND FUTURE CHALLENGES

The PDB Core Archive continues to grow year-on-year (Figure 1A), while incoming structures grow in size and complexity (Figure 4). The wwPDB partners have addressed these challenges by launching and continually improving the unified OneDep deposition–validation–biocuration system. Global adoption of best practices and increased automation have improved the Data Depositor experience and made wwPDB biocuration more efficient.

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Instead of devoting efforts to unnecessary repetitious work on relatively uncomplicated depositions, our biocurators are now able to focus more of their time on the more complex depositions, thereby improving data consistency and accuracy. ORCiD IDs (orcid.org) for OneDep Depositor(s) of Record were made mandatory in 2018. To further improve the depositor experience and enable better management of incoming data, OneDep protocols will be changed to allow login using ORCiD persistent digital identifier unique to each researcher.

Rapidly evolving experimental methods, such as SFX/XFEL and 3DEM, require frequent extensions of the PDBx/mmCIF data dictionary for expanded collection of related metadata using the OneDep system. Validation of 3DEM structural data also requires further development of OneDep, particularly for validation of 3D atomic coordinates against Coulomb potential maps that will allow better assessment of 3DEM structure quality. Emerging methods, such as Integrative/Hybrid Methods (I/HM), present entirely new challenges in data representation and validation. In 2014, the wwPDB established a wwPDB Hybrid Methods Task Force, which produced a white paper detailing the outcome of its inaugural meeting (39). These recommendations led to development of a prototype system (PDB-Dev, pdb-dev.wwpdb.org) for representation, deposition, and archiving of I/HM structure data (41, 42). In parallel, as a first step towards inclusion of I/HM structures in the PDB Core Archive, the OneDep system, in partnership with SASBDB (12), recently introduced combined data deposition for structures determined using SAS data in addition to the use of traditional structure determination techniques.

To avoid fragmentation of structural biology data in different archives that do not interoperate with one another, wwPDB partners are leading efforts to coordinate archiving activities across the discipline. Going beyond the two Core Archives, the PDBx/mmCIF data dictionary contains data items that provide links for individual 3D structures with related data stored by other specialist data resources, such as SBGRID (10), IIRMC (11) and EMPIAR (13).

PLANS FORWARD

The wwPDB is committed to ensuring that all data in the PDB Core archive are as accurate and consistent as possible. Three major initiatives are planned for the coming years.

First, Depositors of Record will be able to make corrections to existing structures in the PDB Core Archive by updating the atomic coordinates, while preserving the original PDB identifier to improve ligand structures or make a better quality structure available for a particular macromolecule. The recent introduction of versioning makes this long-desired opportunity feasible for the first time.

Second, work is underway to further enhance the original wwPDB validation report. Of particular importance will be enhanced validation for both NMR and 3DEM structures. For NMR, the OneDep system will restrict deposition of restraints to the NEF (40) or NMR-STAR (25, 26) formats, and the future validation reports will include analysis of NMR experimental restraint data. The archival format for NMR restraint data will continue to be NMR-STAR. For 3DEM, the wwPDB partners are working with EMDB to improve validation of atomic models built using Coulomb potential maps. Finally, ligand representation and validation will be improved as recommended in the Ligand Validation Workshop white paper (31). All of this work will be informed by ongoing discussions with the various wwPDB Validation Task Forces and other community experts.

Third, the wwPDB plans to develop a new mechanism to resolve the official DOI for each PDB structure. Evolution of the wwPDB PDB Core Archives has resulted in there being multiple data files associated with a given 3D structure, including the atomic coordinates, experimental data, validation reports, and other associated files. The wwPDB partners plan to introduce a new wwPDB web page accessible from the official DOI that will provide access to all relevant files across the two Core Archives. The wwPDB strongly encourages all scientific journals to link to these pages using the DOI for each newly published 3D structure once these pages are made available.

HISTORICAL PERSPECTIVE AND POSTSCRIPT

The PDB has undergone enormous changes since its humble beginnings with just seven structures in 1971. Notwithstanding seismic shifts in the discipline that we now call structural biology and 20,000-fold growth in the PDB, management of the resource as the single global archive of 3D structure data for biological macromolecules continues to be underpinned by an unwavering commitment to universal access to high data quality without limitations on usage. Initially, what is now referred to as the PDB Core Archive was managed entirely within the United States, first at Brookhaven National Laboratory (1), and then by the RCSB PDB, a consortium formed by Rutgers University, the San Diego Supercomputer Center/University of California San Diego and the National Institute of Standards and Technology (3). Since 1999, PDBj and PDBe (formerly known as Macromolecular Structure Database (MSD)) informally worked with RCSB PDB to support PDB deposition and biocuration. In 2003, this arrangement was formalized, and joint international management of the resource began with founding of the wwPDB by the RCSB PDB, PDBe, and PDBj (2). BMRB joined the wwPDB in 2006.

Above all since 2003, joint management by the wwPDB has ensured that the resource remains the single global archive for 3D macromolecular structure data, becoming a central player in the international biological data ecosystem. In addition, joint management has enabled a host of important accomplishments, including (i) adoption of the PDBx/mmCIF data dictionary, (ii) multiple rounds of archive-wide remediation to improve data consistency and data quality, (iii) mandatory deposition of experimental data, (iv) development of community standards for validation of structures and related data/metadata from multiple experimental methods, (v) launch of the universal OneDep deposition/validation/biocuration system and (vi) launch of the PDB-Dev prototype for archiving I/HM structures.

As in 2003, the wwPDB partners remain committed to working together with their diverse User communities to confront myriad challenges presented by ever more complex structures and related data/metadata. Many of the

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

most exciting structures entering the PDB Core Archive today are determined using the rapidly evolving techniques of SFX/XFEL and 3DEM. The I/HM structures on the horizon promise to be even more important contributors to research and education in biomedicine as 3D structures of macromolecular machines at work inside cells come from cryo-electron tomography combined with other methods. Members of the wwPDB organization look forward to addressing these challenges and ensuring that joint management of the wwPDB Core Archives continues to serve Users around the globe.

FUNDING

The RCSB Protein Data Bank is jointly funded by the National Science Foundation; National Institute of General Medical Sciences; National Cancer Institute; Department of Energy [NSF-DBI 1338415]. The Protein Data Bank in Europe is supported by European Molecular Biology Laboratory-European Bioinformatics Institute; Wellcome Trust [88944, 104948]; Biotechnology and Biological Sciences Research Council [BB/G022577/1, BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1, BB/M011674/1, BB/M020347/1, BB/M020428/1]; European Union [284209, 675858]. The Protein Data Bank Japan is supported by the Database Integration Coordination Program from the National Bioscience Database Centre (NBDC)-JST (Japan Science and Technology Agency) and the joint usage program of Institute for Protein Research, Osaka University. The BMRB is supported by the U.S. National Institutes of Health [R01GM109046]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Protein Data Bank (1971) Protein Data Bank. *Nat. New Biol.*, **233**, 223.
2. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980–980.
3. Berman,H.M., Westbook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
4. Mir,S., Alhroub,Y., Anyango,S., Armstrong,D.R., Berrisford,J.M., Clark,A.R., Conroy,M.J., Dana,J.M., Deshpande,M., Gupta,D. *et al.* (2017) PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.*, **46**, D486–D492.
5. Kinjo,A.R., Bekker,G.-J., Suzuki,H., Tsuchiya,Y., Kawabata,T., Ikegawa,Y. and Nakamura,H. (2017) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.*, **45**, D282–D288.
6. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
7. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
8. Abbott,S., Iudin,A., Korir,P.K., Somasundharam,S. and Patwardhan,A. (2018) EMDB Web Resources. *Curr. Protoc. Bioinforma.*, **61**, 5.10.1–5.10.12.
9. Young,J.Y., Westbrook,J.D., Feng,Z., Sala,R., Peisach,E., Oldfield,T.J., Sen,S., Gutmanas,A., Armstrong,D.R., Berrisford,J.M. *et al.* (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure*, **25**, 536–545.
10. Morin,A., Eisenbraun,B., Key,J., Sanschagrin,P.C., Timony,M.A., Ottaviano,M. and Sliz,P. (2013) Collaboration gets the most out of software. *Elife*, **2**, e01456.
11. Grabowski,M., Langner,K.M., Cymborowski,M., Porebski,P.J., Sroka,P., Zheng,H., Cooper,D.R., Zimmerman,M.D., Elsliger,M.A., Burley,S.K. *et al.* (2016) A public database of macromolecular diffraction experiments. *Acta Crystallogr. D Struct. Biol.*, **72**, 1181–1193.
12. Valentini,E., Kikhney,A.G., Previtali,G., Jeffries,C.M.C.M. and Svergun,D.I. (2015) SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.*, **43**, D357–D363.
13. Iudin,A., Korir,P.K., Salavert-Torres,J., Kleywegt,G.J. and Patwardhan,A. (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, **13**, 387–388.
14. Gore,S., Velankar,S. and Kleywegt,G.J. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **68**, 478–483.
15. Gore,S., Sanz García,E., Hendrickx,P.M.S., Gutmanas,A., Westbrook,J.D., Yang,H., Feng,Z., Baskaran,K., Berrisford,J.M., Hudson,B.P. *et al.* (2017) Validation of structures in the Protein Data Bank. *Structure*, **25**, 1916–1927.
16. Henderson,R., Sali,A., Baker,M.L., Carragher,B., Devkota,B., Downing,K.H., Egelman,E.H., Feng,Z., Frank,J., Grigorieff,N. *et al.* (2012) Outcome of the first electron microscopy validation task force meeting. *Structure*, **20**, 205–214.
17. Montelione,G.T., Nilges,M., Bax,A., Güntert,P., Herrmann,T., Richardson,J.S., Schwieters,C.D., Vranken,W.F., Vuister,G.W., Wishart,D.S. *et al.* (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure*, **21**, 1563–1570.
18. Adams,P.D., Aertgeerts,K., Bauer,C., Bell,J.A., Berman,H.M., Bhat,T.N., Blaney,J.M., Bolton,E., Bricogne,G., Brown,D. *et al.* (2016) Outcome of the first wwPDB/CCDC/D3R ligand validation workshop. *Structure*, **24**, 502–508.
19. Read,R.J., Adams,P.D., Arendall,W.B., Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lütteke,T., Otwinowski,Z. *et al.* (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure*, **19**, 1395–1412.
20. Westbrook,J.D., Shao,C., Feng,Z., Zhuravleva,M., Velankar,S. and Young,J. (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, **31**, 1274–1278.
21. Young,J.Y., Feng,Z., Dimitropoulos,D., Sala,R., Westbrook,J., Zhuravleva,M., Shao,C., Quesada,M., Peisach,E. and Berman,H.M. (2013) Chemical annotation of small and peptide-like molecules at the Protein Data Bank. *Database*, **2013**, bat079.
22. Westbrook,J.D. and Bourne,P.E. (2000) STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics*, **16**, 159–168.
23. Fitzgerald,P.M.D., Westbrook,J.D., Bourne,P.E., McMahon,B., Watenpaugh,K.D. and Berman,H.M. (2005) The Macromolecular dictionary (mmCIF). In: *International Tables for Crystallography, Volume G*. pp. 295–443.
24. Westbrook,J., Henrick,K., Ulrich,E.L. and Berman,H.M. (2005) The Protein Data Bank exchange data dictionary. In: *International Tables for Crystallography, Volume G*. pp. 195–198.
25. Ulrich,E.L., Argentar,D., Klimowicz,A., Westler,W.M. and Markley,J.L. (1996) STAR/CIF macromolecular NMR data dictionaries and data file formats. *Acta Crystallogr. A Found. Crystallogr.*, **52**, C577–C577.
26. Doreleijers,J.F., Mading,S., Maziuk,D., Sojourner,K., Yin,L., Zhu,J., Markley,J.L. and Ulrich,E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR*, **26**, 139–146.
27. Tagari,M., Newman,R., Chagoyen,M., Carazo,J.M. and Henrick,K. (2002) New electron microscopy database and deposition system. *Trends Biochem Sci.*, **27**, 589.
28. Westbrook,J., Ito,N., Nakamura,H., Henrick,K. and Berman,H.M. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.
29. Kinjo,A.R., Suzuki,H., Yamashita,R., Ikegawa,Y., Kudou,T., Igarashi,R., Kengaku,Y., Cho,H., Standley,D.M., Nakagawa,A. *et al.* (2012) Protein Data Bank Japan (PDBj): maintaining a structural

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

data archive and resource description framework format. *Nucleic Acids Res.*, **40**, D453–D460.
30. Young,J.Y., Westbrook,J.D., Feng,Z., Peisach,E., Persikova,I., Sala,R., Sen,S., Berrisford,J.M., Swaminathan,G.J., Oldfield,T.J. *et al.* (2018) Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database (Oxford)*, **2018**, doi:10.1093/database/bay002.
31. Smart,O.S., Horský,V., Gore,S., Svobodová Vařeková,R., Bendová,V., Kleywegt,G.J. and Velankar,S. (2018) Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr. D Struct. Biol.*, **74**, 228–236.
32. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Moult,J., Schwede,T. and Tramontano,A. (2018) Evaluation of the template-based modeling in CASP12. *Proteins Struct. Funct. Bioinforma.*, **86**, 321–334.
33. Lensink,M.F., Velankar,S., Baek,M., Heo,L., Seok,C. and Wodak,S.J. (2018) The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins*, **86**(Suppl. 1), 257–273.
34. Haas,J., Barbato,A., Behringer,D., Studer,G., Roth,S., Bertoni,M., Mostaguir,K., Gumienny,R. and Schwede,T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, **86**(Suppl. 1), 387–398.
35. Gaieb,Z., Liu,S., Gathiaka,S., Chiu,M., Yang,H., Shao,C., Feher,V.A., Walters,W.P., Kuhn,B., Rudolph,M.G. *et al.* (2018) D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.*, **32**, 1–20.
36. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2007) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
37. Groom,C.R. and Allen,F.H. (2014) The Cambridge Structural Database in retrospect and prospect. *Angew. Chem. Int. Ed. Engl.*, **53**, 662–671.
38. Trewhella,J., Hendrickson,W.A., Kleywegt,G.J., Sali,A., Sato,M., Schwede,T., Svergun,D.I., Tainer,J.A., Westbrook,J. and Berman,H.M. (2013) Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. *Structure*, **21**, 875–881.
39. Sali,A., Berman,H.M., Schwede,T., Trewhella,J., Kleywegt,G., Burley,S.K., Markley,J., Nakamura,H., Adams,P., Bonvin,A.M.J.J. *et al.* (2015) Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure*, **23**, 1156–1167.
40. Gutmanas,A., Adams,P.D., Bardiaux,B., Berman,H.M., Case,D.A., Fogh,R.H., Güntert,P., Hendrickx,P.M.S., Herrmann,T., Kleywegt,G.J. *et al.* (2015) NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.*, **22**, 433–434.
41. Vallat,B., Webb,B., Westbrook,J.D., Sali,A. and Berman,H.M. (2018) Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure*, **26**, 894–904.
42. Burley,S.K., Kurisu,G., Markley,J.L., Nakamura,H., Velankar,S., Berman,H.M., Sali,A., Schwede,T. and Trewhella,J. (2017) PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure*, **25**, 1317–1318.

CURRENT WWPDB CONSORTIUM MEMBERS WITH AFFILIATIONS

RCSB Protein Data Bank:

Stephen K. Burley[1,2,3], Helen M. Berman[1], Charmi Bhikadiya[1], Chunxiao Bi[3], Li Chen[1], Luigi Di Costanzo[1], Cole Christie[3], Jose M. Duarte[3], Shuchismita Dutta[1], Zukang Feng[1], Sutapa Ghosh[1], David S. Goodsell[1,4], Rachel Kramer Green[1], Vladimir Guranovic[1], Dmytro Guzenko[2], Brian P. Hudson[1], Yuhe Liang[1], Robert Lowe[1], Ezra Peisach[1], Irina Periskova[1], Chris Randle[3], Alexander Rose[3], Monica Sekharan[1], Chenghua Shao[1], Yi-Ping Tao[1], Yana Valasatava[3], Maria Voigt[1], John Westbrook[1], Jasmine Young[1], Christine Zardecki[1], and Marina Zhuravleva[1]

Protein Data Bank Japan:

Genji Kurisu[5], Haruki Nakamura[5], Yumiko Kengaku[5], Hasumi Cho[5], Junko Sato[5], Ju Yaen Kim[5], Yasuyo Ikegawa[5], Atsushi Nakagawa[5], Reiko Yamashita[5], Takahiro Kudou[5], Gert-Jan Bekker[5], Hirofumi Suzuki[5], Takeshi Iwata[5], Masashi Yokochi[5], Naohiro Kobayashi[5], and Toshimichi Fujiwara[5]

Protein Data Bank in Europe:

Sameer Velankar[6], Gerard J. Kleywegt[6], Stephen Anyango[6], David R. Armstrong[6], John M. Berrisford[6], Matthew J. Conroy[6], Jose M. Dana[6], Mandar Deshpande[6], Paul Gane[6], Romana Gáborová[7], Deepti Gupta[6], Aleksandras Gutmanas[6], Jaroslav Koča[7], Lora Mak[6], Saqib Mir[6], Abhik Mukhopadhyay[6], Nurul Nadzirin[6], Sreenath Nair[6], Ardan Patwardhan[8], Typhaine Paysan-Lafosse[6], Lukas Pravda[6], Osman Salih[8], David Sehnal[6,7], Mihaly Varadi[6], and Radka Vařeková[7]

Biological Magnetic Resonance data Bank

John L. Markley[9], Jeffrey C. Hoch[10], Pedro R. Romero[9], Kumaran Baskaran[9], Dimitri Maziuk[9], Eldon L. Ulrich[9], Jonathan R. Wedell[9], Hongyang Yao[9], Miron Livny[11], Yannis E. Ioannidis[12]

[1] Research Collaboratory for Structural Bioinformatics Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[2] Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA

[3] Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA

[4] The Scripps Research Institute, La Jolla, CA 92037, USA

[5] Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

[6] Protein Data Bank in Europe, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[7] CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamencie 5, 625 00 Brno-Bohuice, Czech Republic

[8] Cellular Structure and 3D Bioimaging, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[9] Biochemistry Department, University of Wisconsin-Madison, Madison, WI 53706, USA

[10] Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT 06030, USA

[11] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

[12] ATHENA Research and Innovation Center, Athens, Greece

Downloaded from academic.oup.com/nar/article/47/D1/D520/5144142 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Nucleic Acids Research, 2026, **54**, D358–D362
https://doi.org/10.1093/nar/gkaf1226
Advance access publication date: 22 November 2025
Database issue

AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage

Damian Bertoni[1], Maxim Tsenkov[1], Paulyna Magana[1], Sreenath Nair[1], Ivanna Pidruchna[1], Marcelo Querino Lima Afonso[1], Adam Midlik[1], Urmila Paramval[1], Dare Lawal[1], Ahsan Tanweer[1], Meera Last[2], Risha Patel[2], Agata Laydon[2], Dariusz Lasecki[2], Nick Dietrich[2], Hamish Tomlinson[2], Augustin Žídek[2], Tim Green[2], Oleg Kovalevskiy[2], Andy Lau[3], Shaun Kandathil[3], Nicola Bordin[4], Ian Sillitoe[4], Milot Mirdita[5], David Jones[3], Christine Orengo[4], Martin Steinegger[5,6,7,8], Jennifer R. Fleming[1], Sameer Velankar[1,*]

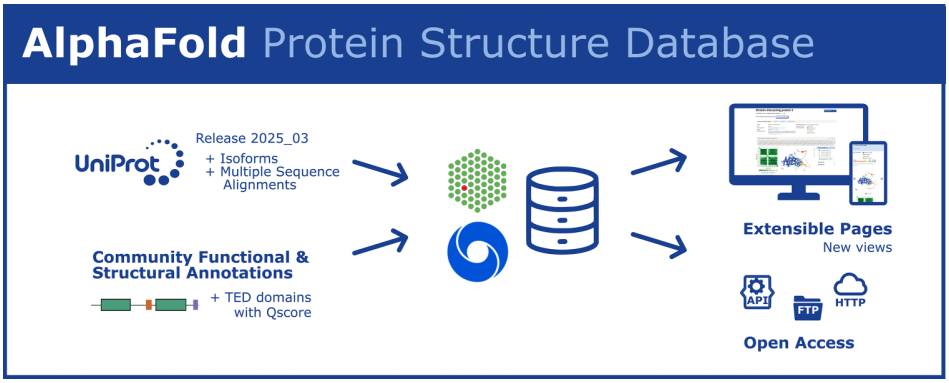
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, CB10 1SD, United Kingdom
[2]Google DeepMind, London, N1C 4DN, United Kingdom
[3]Department of Computer Science, University College London, London, WC1E 6BT, United Kingdom
[4]Institute of Structural and Molecular Biology, University College London, London, England, WC1E 6BT, United Kingdom
[5]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea
[6]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Republic of Korea
[7]Institute of Molecular Biology and Genetics, Seoul 08826, Republic of Korea
[8]Artificial Intelligence Institute, Seoul National University, Seoul 08826, Republic of Korea

*To whom correspondence should be addressed. Email: sameer@ebi.ac.uk

Abstract

The AlphaFold Protein Structure Database (AFDB; https://alphafold.ebi.ac.uk), developed by EMBL–EBI and Google DeepMind, provides open access to hundreds of millions of high-accuracy protein structure predictions, transforming research in structural biology and the wider life sciences. Since its launch, AFDB has become a widely used bioinformatics resource, integrated into major databases, visualization platforms, and analysis pipelines. Here, we report the update of the database to align with the UniProt 2025_03 release, along with a comprehensive redesign of the entry page to enhance usability, accessibility, and structural interpretation. The new design integrates annotations directly with an interactive 3D viewer and introduces dedicated domains and summary tabs. Structural coverage has also been updated to include isoforms plus underlying multiple sequence alignments. Data are available through the website, FTP, Google Cloud, and updated APIs. Together, these advances reinforce AFDB as a sustainable resource for exploring protein sequence–structure relationships.

Graphical abstract



Introduction

The field of life science has been transformed by protein structure prediction tools, such as AlphaFold [1–3], ESMFold [4], and RoseTTAFold [5]. Built upon decades of foundational research [6–8] and open scientific data resources such as the Protein Data Bank (PDB) [9, 10], MGnify [11] and the Univer-

Received: September 30, 2025. **Revised:** October 18, 2025. **Accepted:** October 18, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/54/D1/D358/8340156 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

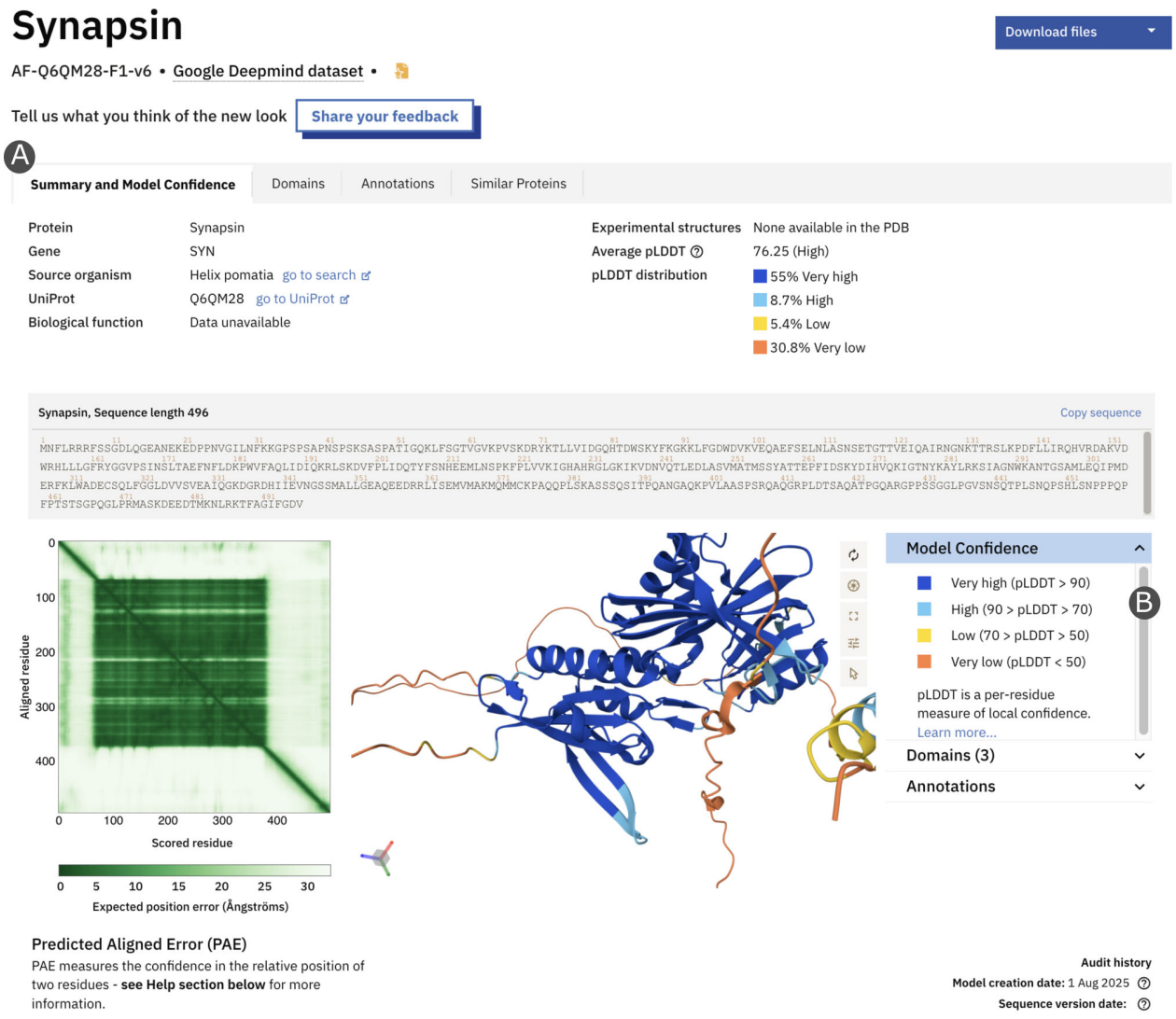


Figure 1. Example of summary page for Synapsin AF-Q6QM28-F1-v6. (**A**) The page is now arranged with tabs separating each section. The central panel displays the predicted 3D model, coloured by per-residue confidence score (pLDDT). (**B**) The right panel accordion provides an overview of domains and annotations and allows the user to toggle between different colour schemes.

sal Protein Resource (UniProt) [12], these advanced methods have dramatically narrowed the gap between the vast number of known protein sequences and the availability of structure models.

The AlphaFold Protein Structure Database (AFDB), a collaboration between Google DeepMind and EMBL–EBI [13–15], was established to democratize access to millions of highly accurate protein structure predictions. The practical utility of these predictions is underscored by their use, with to date, over 4,501,953 total users accessing the AFDB website, more than 18,000 proteome archives downloaded, and the data's integration into major primary data resources [16, 17] and molecular visualization software [18–20]. This widespread adoption has benefited various life science disciplines from structural biology and bioinformatics to drug discovery [21–27].

To meet the ongoing demand for comprehensive structural data, AFDB has continued to evolve with new tools and functionalities [15]. However, synchronization with UniProt fell several years behind, leaving the database increasingly out of date and limiting coverage of newly described proteins [28]. The current release restores alignment with UniProt (2025_03), ensuring that predictions once again reflect the evolving sequence space. Newly generated multiple sequence alignments (MSAs) are provided, and coverage has been extended to include isoform-specific predictions, broadening biological relevance and accuracy. In parallel, a redesigned entry page enhances usability and interpretation.

Enhanced user experience and functionalities

To prioritize an intuitive and accessible user experience, AFDB has now undergone a redesign (Fig. 1). A central element of this overhaul is the enhanced prominence of the structure viewer, now featuring integrated annotations for a more streamlined analysis. New, focused tabs, such as the new Domains tab, are directly linked to the 3D viewer, simplifying complex visualization and analysis by providing immediate

Downloaded from academic.oup.com/nar/article/54/D1/D358/8340156 by Kungl Tekniska Högskolan Biblioteket user on 02 October 2026

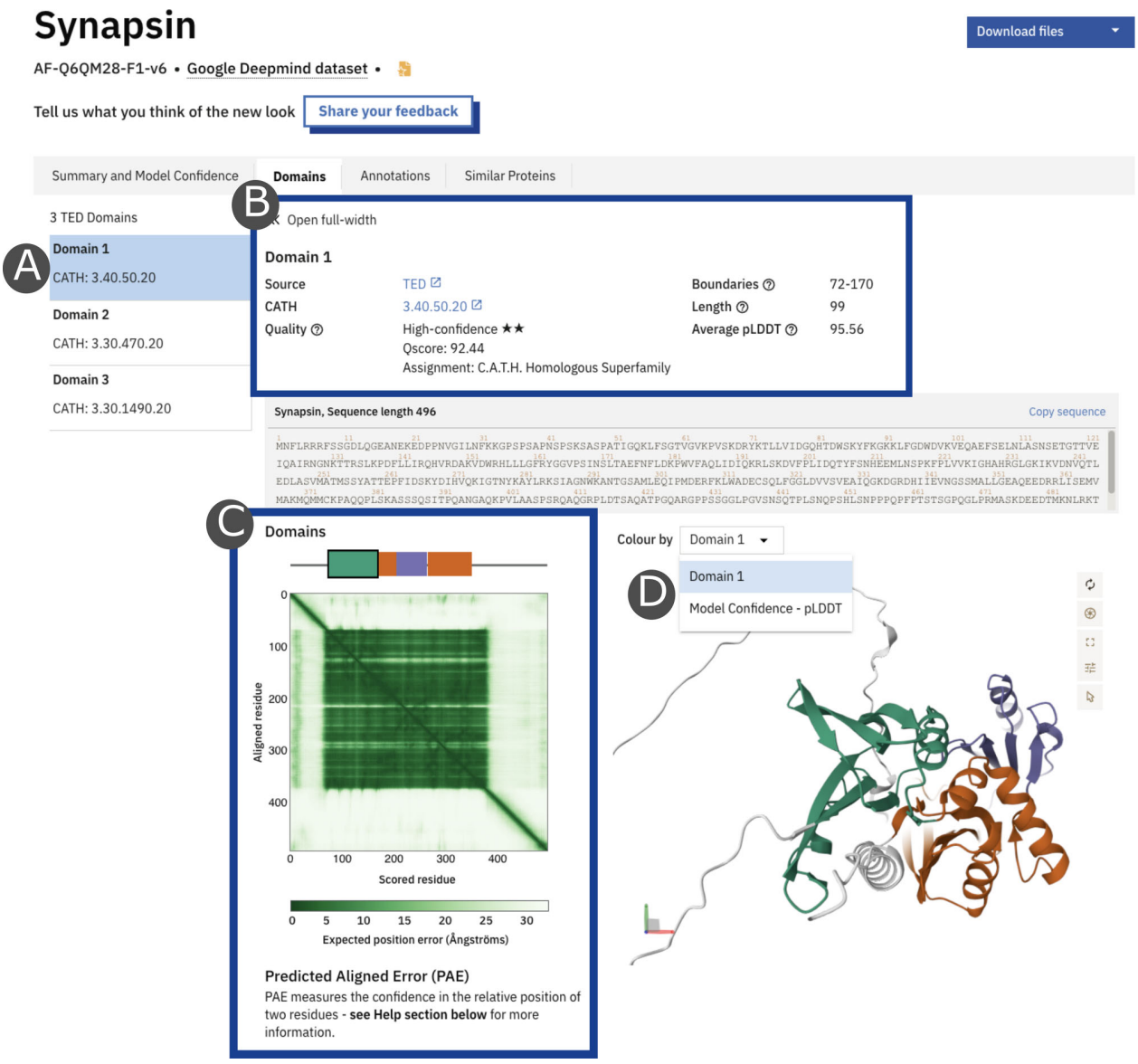


Figure 2. Example of domains tab for Synapsin AF-Q6QM28-F1-v6. This tab provides a comprehensive overview of the information available for a single protein domain provided by TED in the AFDB. (**A**) The panel on the left shows the complete list of domains identified in the entry. (**B**) The descriptive and quality assessment key metrics for a specific domain, including boundaries, length, mean pLDDT, and Qscore, are shown along with the high-confidence CATH assignment. (**C**) The predicted aligned error plot provides a visual representation of the AF2-predicted confidence in the orientation of the domains relative to the rest of the protein. (**D**) Colours on the model can be toggled between domain and pLDDT.

structural insights. This functionality, developed with accessibility and usability principles in mind, lowers the entry barrier for expert and non-expert users, making use of advanced structural biology data more inclusive and impactful for the global scientific community. This commitment to user-centred design ensures that the powerful insights from AlphaFold predictions are readily available and comprehensible to a broad audience.

Data update

The AFDB collection has now been updated to UniProt release (2025_03), incorporating Swiss-Prot/UniProtKB isoforms. The integration of isoforms offers a more complete view of proteome diversity and has revealed widespread changes in protein domain architecture and interaction interfaces [29, 30]. Furthermore, we now provide the underlying MSAs and MSA depths used to predict the structures, offering a new layer of data for evolutionary and structural analyses.

Integrating the Encyclopedia of Domains

To facilitate functional understanding of the proteins in AFDB, over 361 million domain annotations from the Encyclopedia of Domains (TED) [21] covering over 165M proteins have now been added. The TED methodology employs a state-of-the-art, consensus-based approach that integrates three deep-learning-based domain parsing algorithms to robustly determine domain boundaries. These domain assignments are subsequently aligned against the CATH database (Version 4.4)

Downloaded from academic.oup.com/nar/article/54/D1/D358/8340156 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

[31], using structural alignment algorithms. Each domain entry in the AFDB is associated with a suite of descriptive and quality assessment metrics (Fig. 2). A consensus level serves as an initial quality indicator, reflecting the degree of agreement among the three domain-parsing methods. For domains with a structural match, a CATH identifier and assignment are provided based on the level of the hierarchy matched. A second key metric, the Qscore, provides a comprehensive, composite measure of the quality of each domain assignment (details in Supplementary data). This resource supports evolutionary, structural, and functional studies by providing a deeper view of domain organization.

Data availability and API updates

The AFDB prioritizes open data access. All data, including new additions, are accessible via our website (https://alphafold.ebi.ac.uk). The EMBL–EBI's FTP area hosts TAR files for proteomes of 46 organisms, including model organisms and WHO pathogens of interest (ftp.ebi.ac.uk/pub/databases/alphafold). The full AlphaFold dataset is accessible from Google Cloud Public Datasets, with user guidance at https://alphafold.ebi.ac.uk/download and version history documented at https://ftp.ebi.ac.uk/pub/databases/alphafold/CHANGELOG.txt. To support continued development, AFDB is introducing breaking changes to its API schema. The legacy API will be retired in June 2026 (https://alphafold.ebi.ac.uk/api-docs).

Discussion and future directions

The latest updates to AFDB bring the resource back in line with UniProt and introduce an extensible interface designed to support future developments that enrich data and functionality. The inclusion of isoform-specific predictions and newly generated MSAs expands structural coverage and adds essential biological and evolutionary context to the models. This, in turn, lowers the barrier to entry for structure-informed research across the life sciences, enhances reproducibility, and promotes sustainable and greener research practices.

Looking ahead, future releases will continue to work with the scientific community to focus on new datasets and annotations, ensuring that AFDB reflects both the expanding sequence space and the evolving needs of researchers. Its development is guided by three principles: (i) filling gaps in structural coverage, (ii) improving existing models to enhance accuracy and utility, and (iii) working with the scientific community to address global challenges, from antimicrobial resistance to food security. A key priority is the inclusion of structural models and their validation in ways that maximize their scientific impact. Central to this vision is fostering community participation in annotation, building on the successful PDBe-KB model of collaborative consortia. By coupling structural predictions with rich, community-contributed functional insights, AFDB will continue to evolve into a knowledge-rich resource that accelerates discovery and amplifies the impact of protein structure data.

Acknowledgements

We acknowledge the foundational contributions of researchers who consistently deposit structures into public databases, thereby enriching the field of structural biology. We are grateful for the open data provided by major repositories, including UniProt, MGnify, and the PDB, which are indispensable for the training of structure prediction tools. Lastly, we thank the PDBe and Google DeepMind teams for their rigorous and critical evaluation of new features; their collective contributions have been highly beneficial to the development of this work.

Author Contributions: Damian Bertoni (Data curation [equal], Formal analysis [equal], Methodology [equal], Software [lead], Validation [equal]), Maxim Tsenkov (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [supporting], Writing—original draft [equal], Writing—review & editing [supporting]), Paulyna Magana (Writing—original draft [equal], Writing—review & editing [supporting]), Sreenath Nair (Data curation [supporting], Methodology [supporting], Project administration [supporting], Software [equal], Supervision [equal], Validation [equal]), Ivanna Pidruchna (Validation [equal], Visualization [lead]), Marcelo Querino Lima Afonso (Software [equal], Validation [supporting]), Adam Midlik (Methodology [supporting], Software [supporting], Validation [supporting], Visualization [supporting]), Urmila Paramval (Software [lead], Validation [supporting], Visualization [supporting]), Dare Lawal (Software [supporting], Validation [supporting]), Ahsan Tanweer (Software [equal]), Meera Last (Project administration [supporting], Resources [supporting]), Risha Patel (Project administration [supporting], Writing—review & editing [supporting]), Agata Laydon (Resources [equal], Supervision [equal]), Dariusz Lasecki (Software [equal]), Nick Dietrich (Software [equal]), Hamish Tomlinson (Software [equal]), Augustin Žídek (Data curation [equal], Methodology [equal], Software [equal], Supervision [equal], Validation [equal], Writing—review & editing [supporting]), Tim Green (Validation [equal]), Oleg Kovalevskiy (Supervision [equal], Validation [equal], Writing—review & editing [equal]), Andy Lau (Software [equal]), Shaun Kandathil (Software [equal], Validation [equal]), Nicola Bordin (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal]), Ian Sillitoe (Software [equal]), Milot Mirdita (Formal analysis [equal], Software [equal], Validation [equal]), David Jones (Conceptualization [equal], Formal analysis [equal], Funding acquisition [equal], Project administration [equal], Software [equal], Supervision [equal]), Christine Orengo (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Resources [equal], Supervision [equal]), Martin Steinegger (Conceptualization [equal], Funding acquisition [equal], Resources [equal], Software [equal], Supervision [equal]), Jennifer R. Fleming (Funding acquisition [equal], Resources [equal], Software [equal], Supervision [equal] Conceptualization [supporting], Project administration [equal], Supervision [equal], Validation [equal], Writing—original draft [supporting], Writing—review & editing [equal]), and Sameer Velankar (Conceptualization [equal], Funding acquisition [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Downloaded from academic.oup.com/nar/article/54/D1/D358/8340156 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Funding

This work was supported by the BBRSC (20-BBSRC/NSF-BIO and BB/Y000455/1 to S.V.; BB/W018802/1 to C.O.; BB/T019409/1 to A.M.L. and D.T.J.; and BB/W008556/1 to S.M.K. and D.T.J.) and the Welcome Trust (221327/Z/20/Z to C.O. and 310300/Z/24/Z to S.V.). M.S. acknowledges support by the National Research Foundation of Korea grants (RS-2020-NR049543, RS-2021-NR061659 and RS-2021-NR056571, RS-2024-00396026) and Creative-Pioneering Researchers Program and Novo Nordisk Foundation (NNF24SA0092560). M.M. acknowledges support from the National Research Foundation of Korea (RS-2023-00250470). Google DeepMind and the European Molecular Biology Laboratory fund the AlphaFold Protein Structure Database. Funding to pay the Open Access publication charges for this article was provided by European Molecular Biology Laboratory.

Data availability

The AFDB is available at https://alphafold.ebi.ac.uk.

References

1. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2
2. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;630:493–500. https://doi.org/10.1038/s41586-024-07487-w
3. Evans R, O'Neill M, Pritzel A *et al.* Protein complex prediction with AlphaFold-Multimer. bioRxiv, https://doi.org/10.1101/2021.10.04.463034, 10 March 2022, preprint: not peer reviewed.
4. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123–30. https://doi.org/10.1126/science.ade2574
5. Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;373:871–6. https://doi.org/10.1126/science.abj8754
6. Yanofsky C, Horn V, Thorpe D. Protein structure relationships revealed by mutational analysis. *Science* 1964;146:1593–4. https://doi.org/10.1126/science.146.3651.1593
7. Göbel U, Sander C, Schneider R *et al.* Correlated mutations and residue contacts in proteins. *Proteins* 1994;18:309–17.
8. Marks DS, Colwell LJ, Sheridan R *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;6:e28766. https://doi.org/10.1371/journal.pone.0028766
9. Gutmanas A, Alhroub Y, Battle GM *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 2014;42:D285–91. https://doi.org/10.1093/nar/gkt1180
10. consortium wPDB. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;47:D520–8. https://doi.org/10.1093/nar/gky949
11. Richardson L, Allen B, Baldi G *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;51:D753–9. https://doi.org/10.1093/nar/gkac1080
12. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2025;53:D609–17. https://doi.org/10.1093/nar/gkae1010
13. Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;50:D439–44. https://doi.org/10.1093/nar/gkab1061
14. Varadi M, Bertoni D, Magana P *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;52:D368–75. https://doi.org/10.1093/nar/gkad1011
15. Fleming J, Magana P, Nair S *et al.* AlphaFold Protein Structure Database and 3D-Beacons: new data and capabilities. *J Mol Biol* 2025;437:168967. https://doi.org/10.1016/j.jmb.2025.168967
16. Piovesan D, Del Conte A, Mehdiabadi M *et al.* MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Res* 2025;53:D495–503. https://doi.org/10.1093/nar/gkae969
17. Blum M, Andreeva A, Florentino LC *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* 2025;53:D444–56. https://doi.org/10.1093/nar/gkae1082
18. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 2004;60:2126–32. https://doi.org/10.1107/S0907444904019158
19. Meng EC, Goddard TD, Pettersen EF *et al.* UCSF ChimeraX: tools for structure building and analysis. *Protein Science* 2023;32:e4792. https://doi.org/10.1002/pro.4792
20. Waterhouse AM, Procter JB, Martin DMA *et al.* Jalview Version 2—multiple sequence alignment editor and analysis workbench. *Bioinforma Oxf Engl* 2009;25:1189–91. https://doi.org/10.1093/bioinformatics/btp033
21. Barbarin-Bocahu I, Graille M. The X-ray crystallography phase problem solved thanks to *AlphaFold* and *RoseTTAFold* models: a case-study report. *Acta Crystallogr D Struct Biol* 2022;78:517–31. https://doi.org/10.1107/S2059798322002157
22. Terwilliger TC, Afonine PV, Liebschner D *et al.* Accelerating crystal structure determination with iterative *AlphaFold* prediction. *Acta Crystallogr D Struct Biol* 2023;79:234–44. https://doi.org/10.1107/S205979832300102X
23. Chojnowski G. Sequence-assignment validation in cryo-EM models with *checkMySequence*. *Acta Crystallogr D Struct Biol* 2022;78:806–16. https://doi.org/10.1107/S2059798322005009
24. Barrio-Hernandez I, Yeo J, Jänes J *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* 2023;622:637–45. https://doi.org/10.1038/s41586-023-06510-w
25. Lau AM, Bordin N, Kandathil SM *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* 2024;386:eadq4946. https://doi.org/10.1126/science.adq4946
26. Romasanta AK, Wareham J, Pujol Priego L. The impact of research data infrastructures: the case of the AlphaFold database. *CERN IdeaSquare J. Exp. Innov.* 2025;9:42–8.
27. Kovalevskiy O, Mateos-Garcia J, Tunyasuvunakool K. AlphaFold two years on: validation and impact. *Proc Natl Acad Sci USA* 2024;121:e2315002121. https://doi.org/10.1073/pnas.2315002121
28. Tsitsa I, Conev A, David A *et al.* The AlphaFold database ages. bioRxiv, https://doi.org/10.1101/2025.06.22.660930, 27 Spetember 2025, preprint: not peer-reviewed.
29. Song Y, Zhang C, Omenn GS *et al.* Predicting the structural impact of human alternative splicing. *Genome Biol* 2025;26:283. https://doi.org/10.1186/s13059-025-03744-x
30. Yang Y, Xie Y, Li Z *et al.* Systematic characterization of protein structural features of alternative splicing isoforms using AlphaFold 2. bioRxiv, https://doi.org/10.1101/2024.01.30.578053, 19 February 2024, preprint: not peer-reviewed.
31. Orengo C, Michie A, Jones S *et al.* CATH—a hierarchic classification of protein domain structures. *Structure* 1997;5:1093–109. https://doi.org/10.1016/S0969-2126(97)00260-8

Received: September 30, 2025. **Revised:** October 18, 2025. **Accepted:** October 18, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/54/D1/D358/8340156 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Nucleic Acids Research, 2025, **53**, D730–D737
https://doi.org/10.1093/nar/gkae1113
Advance access publication date: 18 November 2024
Database issue

The STRING database in 2025: protein networks with directionality of regulation

Damian Szklarczyk[1,2], Katerina Nastou[3], Mikaela Koutrouli[3], Rebecca Kirsch[3], Farrokh Mehryary[4], Radja Hachilif[1,2], Dewei Hu[3], Matteo E. Peluso[1,2], Qingyao Huang[1,2], Tao Fang[1,2], Nadezhda T. Doncheva[3], Sampo Pyysalo[4], Peer Bork[5,6,7,*], Lars J. Jensen[3,*] and Christian von Mering[1,2,*]

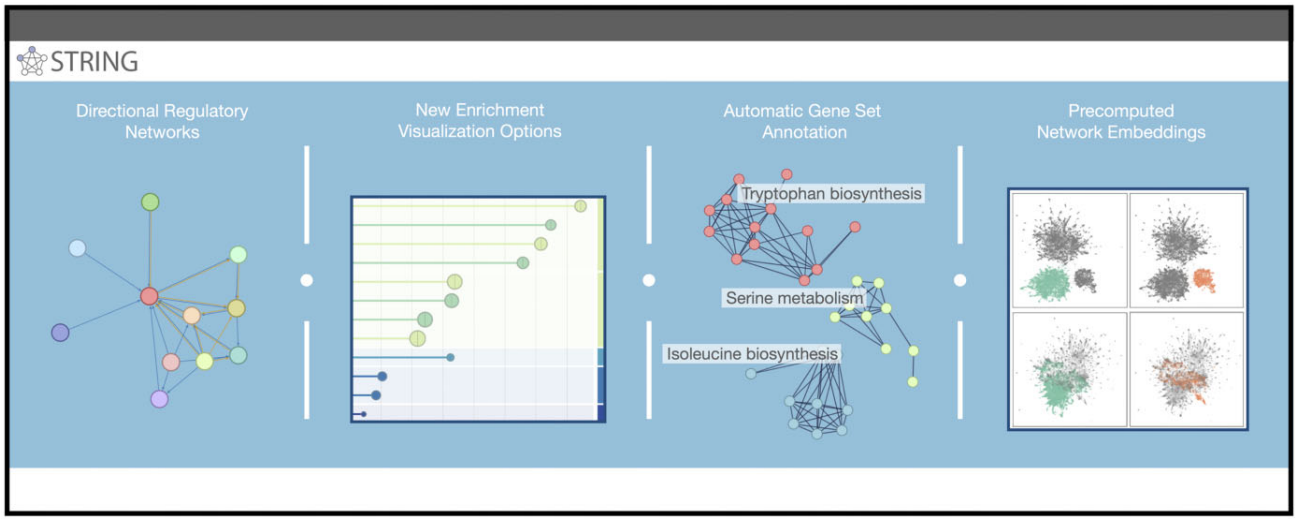
[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, 1015 Lausanne, Switzerland
[3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark
[4]TurkuNLP Lab, Department of Computing, University of Turku, Vesilinnantie 5, 20014 Turku, Finland
[5]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[6]Max Delbrück Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany
[7]Department of Bioinformatics, Biozentrum, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de
Correspondence may also be addressed to Lars J. Jensen. Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Christian von Mering. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch

Abstract

Proteins cooperate, regulate and bind each other to achieve their functions. Understanding the complex network of their interactions is essential for a systems-level description of cellular processes. The STRING database compiles, scores and integrates protein–protein association information drawn from experimental assays, computational predictions and prior knowledge. Its goal is to create comprehensive and objective global networks that encompass both physical and functional interactions. Additionally, STRING provides supplementary tools such as network clustering and pathway enrichment analysis. The latest version, STRING 12.5, introduces a new 'regulatory network', for which it gathers evidence on the type and directionality of interactions using curated pathway databases and a fine-tuned language model parsing the literature. This update enables users to visualize and access three distinct network types—functional, physical and regulatory—separately, each applicable to distinct research needs. In addition, the pathway enrichment detection functionality has been updated, with better false discovery rate corrections, redundancy filtering and improved visual displays. The resource now also offers improved annotations of clustered networks and provides users with downloadable network embeddings, which facilitate the use of STRING networks in machine learning and allow cross-species transfer of protein information. The STRING database is available online at https://string-db.org/.

Graphical abstract



Received: September 15, 2024. **Revised:** October 18, 2024. **Editorial Decision:** October 21, 2024. **Accepted:** October 29, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Introduction

The function of every living cell is primarily governed by a complex network of interacting proteins, with each protein's role determined not only by its molecular activities but also by its position within this network (1,2). Connected proteins work together to contribute to common biological processes through various interaction types, such as physical binding, genetic interactions and regulatory influences. These interactions can collectively be categorized as *functional associations*, which serve as fundamental operational units within biological systems. Unraveling protein networks in their various modalities remains a significant research focus. Consequently, numerous databases have been developed over the years to meet specific research needs. These range from carefully curated pathway databases such as Reactome (3) and KEGG (4), to databases of experimental interaction evidence curated from literature such as BioGRID (5), IntAct (6) and MINT (7), and to those centered on specific modes of interactions, including Complex Portal for protein co-complexes (8) and SIGNOR for regulatory interactions (9). Finally, composite databases such as STRING (10), GeneMANIA (11), FunCoup (12) and HumanNet (13) not only incorporate data from these sources but also employ an array of computational methods to predict additional associations, striving to provide the most comprehensive views of the interactome.

Among these databases, STRING is notable for its many sources of evidence, its robust scoring system, user-friendly interface and comprehensive suite of enrichment features. It is dedicated to assembling a broad set of associations among proteins for the complete proteomes of thousands of organisms across all domains of life. STRING charts interactions ranging from highly confident, well-documented associations to more speculative ones, which are crucial for exploratory and computational analyses. The scoring system ensures that data from diverse sources—including automatic text mining, high- and low-throughput experimental data, and computational predictions—are directly comparable and weighted consistently, regardless of their origin. Additionally, STRING extends and maps networks across species by predicting interologs, thereby broadening the scope of functional associations to encompass a diverse range of organisms. This includes uncultured novel bacterial species derived from metagenomic samples, and even unpublished proteomes uploaded by users.

The development of interaction databases has so far been a trade-off between detailing the exact mode of an interaction—whether physical or functional, signaling or structural, stable or transient—and maintaining a comprehensive set of interactions. Given the limitations of the available data and methodologies, STRING has historically focused on broadly defined functional associations, which provided the most useful set of interactions for proteome-wide analysis. While these networks continue to perform well in downstream tasks (14,15), they often lack fine-grained resolution, leaving more specialized databases better equipped to offer detailed insights into specific interactions if needed. However, the growing volume of data and advancements in text-mining technologies have since enabled the composite, general-purpose databases to catch up, offering more detailed maps of the interaction space. Recently, the STRING database introduced a co-complex (physical) interaction network mode, which details multi-protein assemblies. With the latest update, the database now also includes the largest set of regulatory (directed) interactions, highlighting the flow of information within cells. This major enhancement greatly expands the analytical capabilities of the database, enabling new types of studies that were not previously possible with nondirected interaction networks.

Interaction and pathway databases are frequently used to gain insights into the functional context of individual proteins or to understand the biological organization of an entire protein dataset. Many of these databases enhance the interpretation of protein datasets beyond simple interaction annotations by incorporating an enrichment analysis toolset. This type of analysis involves comparing observed data against expected distributions, enabling researchers to identify statistically significant features or patterns. Enrichment tools typically utilize established ontologies and annotations, such as Gene Ontology (16), MSigDB hallmark sets (17) or OMIM (18). A few databases also leverage their annotated datasets to help understand the functional context of a user's input. For example, KEGG (4) employs pathway and module abstractions to organize genes into datasets for overrepresentation analysis, while Reactome maps interconnected proteins into hierarchically nested modules. Uniquely, STRING uses an unsupervised procedure to hierarchically precluster its entire proteome-wide networks into functional modules. Incorporating these network-derived gene sets into the enrichment analysis facilitates the identification of novel modules, especially in areas of the proteome where manual curation has been less comprehensive so far.

Database content

The basic interaction scope in STRING is that of a 'functional association' between pairs of proteins. A functional association is defined as a contribution of two non-identical proteins to a common function (19,20). This can take many forms; functionally associated proteins can be in physical proximity to each other, regulate each other, exhibit genetic epistasis or even work antagonistically (as long as this occurs in the context of a common function). For the purpose of defining functional associations, the concept of a common function is crucial, but difficult to define—it should broadly be thought of as corresponding to the notion of a 'pathway' or 'function module'. In practice, STRING roughly follows the functional granularity of the pathway maps in the KEGG database (4).

From the set of all functional associations in STRING, subsets are derived that are more specifically annotated with regard to their mechanism of association. Currently, two such more specific subsets are implemented: 'physical' and 'regulatory'. The physical mode refers to pairs of proteins that either bind directly or are at least subunits of the same complex (21). The regulatory mode refers to associated protein pairs that are known to regulate each other's activity in at least one direction. This mode is described in more detail further below; for the first time in STRING, such connections are annotated not only with confidence scores but also with a directionality.

All protein–protein associations in STRING are annotated with 'confidence scores'. These scores are fully precomputed; they scale between 0 and 1 and describe the estimated likelihood of a postulated association being correct, given the available evidence. Separate confidence scores are provided for physical and regulatory modes—these scores provide estimates of the likelihood that the proposed association is taking place and is indeed of the postulated type. The two spe-

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

cific network modes are fully consistent with the functional network, such that if an interaction is present in either the physical or regulatory network, it will, by definition, also be present in the full functional association network (with an equal or higher confidence score). To compute the various confidence scores, the available evidence supporting a given interaction is first separated into 'evidence channels', by the type of evidence. For the basic functional association confidence score, seven such channels are used. These include three that are based on genomic context predictions—neighborhood, fusion and gene co-occurrence—as well as channels dedicated to co-expression, experimental data, curated databases and text mining.

The *genomic context channels* focus on associations inferred from genome sequences alone, detecting evolutionary patterns such as gene proximity, fusion events and co-occurrence across different organisms. For instance, the *neighborhood channel* assigns an association score to gene pairs that are located close to each other on the chromosome in the same orientation (in prokaryotic genomes) (22). The *fusion channel* identifies open reading frames that result from gene fusion events (23), while the *co-occurrence channel* examines whether genes have a nontrivial but shared distribution across genomes (24), implying a shared history of horizontal transfers, losses or duplication events and thus likely related functions. The *co-expression channel* compiles data from gene expression studies, analyzing both transcript and protein abundances across various conditions. By comparing expression profiles, it identifies gene pairs with similar expression patterns, suggesting functional linkage (10,25). The *experiments channel* aggregates interaction evidence from laboratory assays, including biochemical, biophysical and genetic assays. Data are imported from primary repositories such as BioGRID (5) and the IMEx consortium (26), and the confidence scores are estimated by globally benchmarking the accuracy of annotated experimental techniques, as well as within-dataset performance and consistency for the case of high-throughput experiments (10). The *database channel* is based on well-described, curated protein–protein associations from expert-compiled resources, such as KEGG (4), Reactome (3) and Gene Ontology Complexes (16). These resources provide well-established pathways, complexes and functional relationships, offering a high level of reliability. Unlike other channels, the database channel assigns a uniformly high confidence score to associations, reflecting their established nature. Finally, the *text-mining channel* utilizes a large corpus of scientific literature, including PubMed abstracts and full-text articles, to identify co-mentions of protein names (27). By statistically analyzing the frequency of these co-mentions in various textual contexts, this channel uncovers potential associations that may not be evident from structured data alone. In addition, for the more specific association modes, dedicated large language models are employed to detect sentences supporting either physical (21) or regulatory associations (see below).

For each evidence channel, the available interaction evidence is translated into a confidence score by first quantifying the evidence using channel-specific metrics and then converting these metrics into likelihoods using calibration curves derived from prior knowledge (from pathway-map memberships in KEGG). For the more specific association modes 'physical' and 'regulatory', only channels and evidence that are applicable to these modes are considered. After this, all channel scores that have been computed for a given protein pair in a given organism are transferred onto related protein pairs in other organisms, based on the 'interolog' concept (27,28). Lastly, a final, combined confidence score is computed by integrating the channel-specific subscores probabilistically, under the assumption that evidence in different channels is largely independent. Users of STRING can directly rely on this combined score for browsing and filtering networks, or they can alternatively customize their analyses by enabling or disabling specific channels separately, after which the combined score is recomputed according to their chosen settings.

All primary evidence underlying a given interaction can be inspected interactively, in dedicated evidence viewers online. Furthermore, accessory information is available for each protein, such as its three-dimensional structure, domain composition, annotation and cross-references. Apart from its core protein-network functionality, STRING also implements features that allow extensive analysis of uploaded user data. These include functional enrichment detection (29), experimental bias detection (21), homology searches, and clustering and organizing large query protein sets. All data contained in STRING are freely available for download, under a Creative Commons BY 4.0 license. Apart from its website, STRING can also be accessed via a dedicated Cytoscape plugin (30), through an application programming interface (API) (31), as well as through an R/Bioconductor package (https://www.bioconductor.org/packages/release/bioc/html/STRINGdb.html).

Regulatory networks

Molecular networks, like those available through STRING, have become invaluable tools in biomedical research, offering powerful insights into how molecules work together within cells. The functional association networks, while useful for many applications, do not specify the interaction type (e.g. complex formation or transcriptional regulation) nor the direction of interaction (i.e. who regulates whom). Constructing networks with more detailed interaction types and directions can significantly expand their utility, particularly for building mathematical models of biological systems or for making logical inferences. The directionality of interactions is essential for interpretation of omics data, which typically reveals more about downstream effects of conditions such as diseases than it does about the upstream events that led to it and thus could be targeted therapeutically.

To incorporate interactions with specific directionality, sign and type into STRING, we leveraged the novel dataset and deep learning-based relation extraction method described in RegulaTome (32). The RegulaTome corpus provides a rich and diverse dataset that includes 16 961 relations between 54 951 entities annotated across over 2500 documents. This corpus is utilized to train a deep learning-based method used to annotate and extract the regulatory events throughout the entire available literature.

On top of undirected physical interactions—which are already included in STRING—the following types of directed interactions can be extracted using the system developed for RegulaTome: *Regulation*, with the signed subtypes of *Positive Regulation* and *Negative Regulation*, *Regulation of Gene Expression*, *Regulation of Degradation* and *Catalysis of Post-Translational Modifications*, covering six subtypes: *Catalysis of Small Protein Conjugation*, *Catalysis of Small Protein Removal*, *Catalysis of Phosphorylation*, *Catalysis of Dephospho-*

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

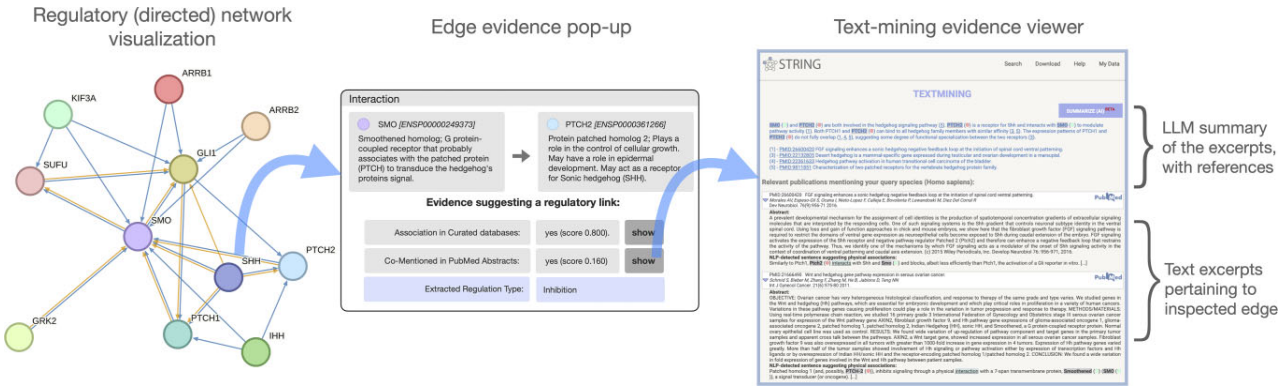


Figure 1. Illustration of the new 'regulatory network' mode in STRING, where the network edges visually indicate the direction, confidence and sources of each regulatory interaction. Clicking on an edge within the network brings up a pop-up window with a detailed overview of the available evidence and the annotated type of regulation. Users can further explore the data behind the interactions by accessing the specific evidence viewers linked within the pop-up, including the text-mining evidence viewer. This viewer presents excerpts from literature pertaining to the inspected interactions, as well as an automatically generated large language model summary of the presented excerpts.

rylation, *Catalysis of Small Molecule Conjugation* and *Catalysis of Small Molecule Removal*. To enhance the interoperability of the annotations, these chosen relation types align with Gene Ontology (32).

To identify and classify a wide array of interaction types, we fine-tuned the RoBERTa-large-PM-M3-Voc language model, a transformer-based model well suited for biological tasks. This model was trained on the RegulaTome dataset for multi-label extraction of the directed, typed and signed interactions mentioned above, and achieved an average F_1 score of 73.5% (with a precision of 75.2% and a recall of 71.8%) on a held-out test set, although the exact performance varied across different types of relationships. This level of accuracy is substantial, considering the complexity of the underlying task. Using this model, we processed over 1.2 billion sentence-level pairs extracted from all available PubMed abstracts and PMC Open Access full-text documents, and assigned 3.5% of them (~43 million, of which ~18 million in human) with at least one positive label indicating directed (72.9%) or signed (33.1%) relationships among the proteins. Furthermore, we made use of the existing curated knowledge of regulatory interactions in the *database* channel, by parsing regulatory information from the SIGNOR, KEGG and Reactome databases.

To integrate these interactions into STRING, we followed a similar score aggregation and benchmarking approach as for physical interactions (21). Here, we perform benchmarking in five categories for which we can derive gold standard datasets of known human regulatory interactions from SIGNOR: *regulation*, *upregulation*, *downregulation*, *transcriptional regulation* and *phosphorylation*. The resulting calibration functions are then applied to extracted regulations of all types to produce the final regulatory confidence scores.

The new regulatory network has been seamlessly integrated into the existing STRING framework, complementing the 'functional' and 'physical' network types. Directional edges within this network are visually represented by arrows, depicting both bidirectional and unidirectional relationships. The network visualization can show varying confidence levels between directions (in the 'confidence' view) or the specific directionality of the sources (in the 'evidence' view). The user interface retains its intuitive design, enabling users to access the type of regulatory event and the evidence for the interaction by clicking on the edge (Figure 1). Additionally, all API functions have been updated to fully support the new network type, which can be accessed by specifying the parameter *network_type=regulatory* in the API call.

Expanded co-expression networks

The *co-expression* channel in STRING identifies genes with similar expression profiles across various tissues and conditions, revealing their involvement in shared biological processes. It compiles data from large-scale gene expression studies, analyzing transcript and protein levels to detect gene pairs with coordinated expression. This method highlights potential functional relationships between genes. Importantly, the analysis is free from study bias, as it directly uses raw omics data instead of relying on manually curated or experimental datasets, providing a more precise and objective view of gene interactions.

For the upcoming version 12.5, we are expanding the *co-expression* channel by generating additional networks through the application of FAVA (Functional Association of Variational Autoencoders) (25) on a wider range of single-cell RNA sequencing data. Specifically, we are incorporating single-cell data from the *cellxgene* Atlas (33) to enhance the human and mouse networks and from the EBI Single Cell Expression Atlas to extend coverage to more model organisms (34). This expansion will provide more detailed, organism-specific co-expression maps, enabling more precise identification of gene co-expression patterns and cross-species comparisons, particularly for development, immune response and disease-related gene expression dynamics.

Improved enrichment analysis

Gene set enrichment analysis is a critical component of high-throughput biological studies, providing key insights into the biological characteristics of datasets. This analysis is pivotal in identifying essential aspects such as biological processes, pathways and phenotypes, which are crucial for understanding the biological implications of experiments. Typically, overrepresentation analysis is employed, wherein a subset of data exceeding a specific cutoff (empirical or customary) is compared against an experimental background. By default, STRING

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

uses the whole genome/proteome background; however, it is recommended that users provide a custom background list representing a more realistic universe of genes/proteins detected by their assay (35). Alternatively, the entire sorted dataset can be analyzed using gene set enrichment analysis to determine whether the distribution of genes at the top, bottom or both ends deviates significantly from what would be expected by chance. STRING supports both of these enrichment methods. The most recent update to STRING introduces various improvements, including a new false discovery rate (FDR) correction, enhanced filtering options and a new graphical visualization.

FDR correction

In its analyses, STRING performs a test for each pathway ('term') within a given pathway collection ('category'). Like many other enrichment tools, STRING employs the Benjamini–Hochberg correction (36) to adjust for multiple hypothesis testing. This statistical correction is essential for managing the FDR effectively. However, the larger the term count and the greater the diversity of the terms, the higher the required correction and the lower the chance of exceeding the alpha level (significance threshold) for a given term. One method used to address this issue involves restricting the analysis to a subset of terms, such as 'GO slims' that focus on broad, high-level terms with a simplified ontology structure (16). However, this approach has some disadvantages: it might exclude terms particularly relevant to a tested gene set, it may not cover all functions by design and it is mainly applicable to hierarchical classification systems such as the Gene Ontology.

STRING utilizes several hierarchical ontologies, such as the Brenda Tissue Ontology, Disease Ontology, Human Phenotype Ontology and the aforementioned Gene Ontology; they typically contain many more leaf-ward nodes than nodes closer to the root. The parental nodes must contain all the genes included in the child terms, creating an imbalance in the ontology with many more smaller terms and fewer larger terms. Removing smaller terms can substantially increase statistical power. However, simply eliminating all small terms could negatively impact discoverability, especially for small query gene sets where the smaller terms might be the most relevant for biological interpretation. The ontologies are usually large, with >10 000 terms (and therefore tests) in the Biological Process branch of the Gene Ontology tree. Such strong corrections have the effect that for larger query gene sets, or for backgrounds that cover only part of the proteome, it might not be statistically possible to detect enrichments for a subset of the terms; these can therefore be removed from consideration.

In its new version, STRING takes a more flexible approach—by only testing terms that have *a priori* any statistical potential for enrichment. This decision is based on several parameters: the size of the term, the size of the query set, the background size and the number of tests conducted. By calculating the smallest and largest term sizes that could theoretically be enriched given the FDR correction, STRING can determine which terms are viable for inclusion in the analysis. Terms outside these bounds are excluded from testing. This methodical exclusion significantly enhances the statistical power of the enrichment analysis, without omitting any terms that could be enriched. For smaller query sets, this strategy will maintain all terms, as even the smallest tested sets (term size = 2) might still be theoretically enriched. However, for larger, less specific sets or for tests against a custom smaller background, this approach markedly improves the statistical power, enabling STRING to perform more focused analyses without the drawbacks of increased false positive rates or omitting potentially biologically interesting terms. Although this method leverages the characteristics of ontologies, it does not require a hierarchy (parent–child relationships between the terms) to function and can be applied to any pathway/term collection with an imbalance of small terms, such as PMID-derived gene sets and STRING neighborhood clusters.

Term filtering

One key insight from our online user surveys is that the often overwhelming number of enrichment results can make it challenging to discern the most relevant enriched terms. We have thus enhanced the display of enrichment results, allowing users to filter and sort them based on (i) FDR, (ii) strength, (iii) signal, (iv) term size and (v) term similarity.

The newest addition to our filtering options are the 'term similarity' and 'signal' filters. The 'term similarity' filter uses the Jaccard index to measure the similarity between the gene sets of terms within the same category. The method proceeds by sorting the terms according to their *P*-values, prioritizing those with the strongest statistical significance. The algorithm then iterates through the list of terms; any term that exhibits a similarity exceeding a predetermined, user-defined threshold relative to terms already included in the results is subsequently excluded. This approach requires no additional information beyond what is already available in the dataset and can be uniformly applied across all categories of terms. The 'signal' filter is defined as a weighted harmonic mean between the ratio of observed to expected gene occurrences in an enriched term and its −log(FDR), respectively. This filter balances the impact of both metrics and provides a more intuitive ordering of enriched terms and has been implemented as the default sorting criterion.

These filters are designed to exclude potentially less relevant results, such as terms that are marginally significant, small or show significant overlap in terms of gene content. This enhancement streamlines the search for relevant observations, helping users focus on the most meaningful results.

Graphical visualization

In response to user feedback requesting more intuitive, graphical visualizations of the detected enrichments, STRING now has adopted an interactive dot plot as a primary display method for enrichment analysis outcomes. We chose the dot plot for its clarity in representing enriched functions across three dimensions: (i) enrichment signal along the *X*-axis, (ii) the FDR indicated through the color coding of the dots and (iii) the term's protein count in the network represented by the size of each dot. The terms (listed on the *Y*-axis) are ranked by their enrichment signal or, if grouped by similarity, by the enrichment signal of their group first (Figure 2). This visualization reflects all filter settings chosen by the user to manage the volume of results displayed. These plots can be additionally customized in the interface and the resulting graphs can be downloaded as PNG for immediate use or as SVG for further modifications.

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

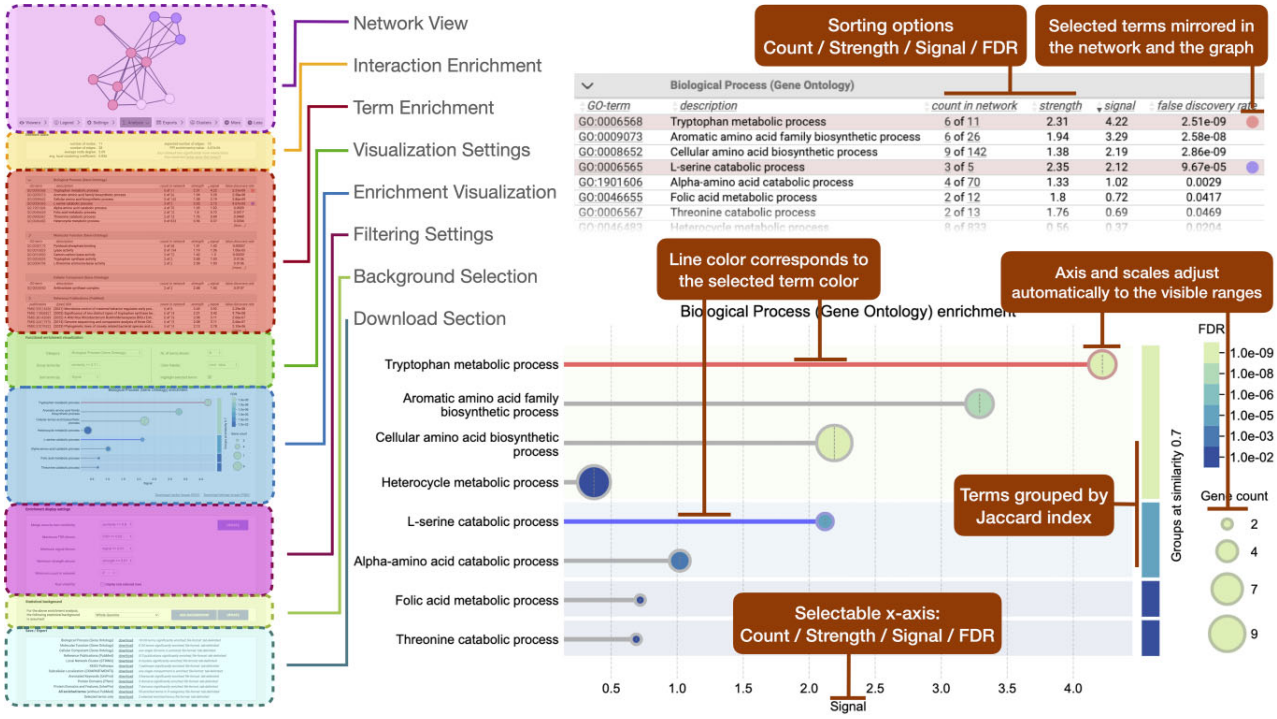


Figure 2. Enrichment analysis interface from the STRING database website. Left: Schematic of the enrichment analysis tab highlighting various sections of the webpage. Top right: A zoomed-in view of the analysis table with two user-highlighted terms (colored lines). Bottom right: A visualization of the enrichment depicting the two terms highlighted in corresponding colors.

Furthermore, a distinctive feature of STRING's visualization is the incorporation of similarity groups, which visually group related terms on the plot, adding an additional dimension to the enrichment results. These groups utilize computed similarities to cluster terms together, highlighting relationships between the terms and aiding the analysis. The clustering employs the average linkage method based on the Jaccard index, with a user-specified cutoff applied to form groups that, on average, share the specified similarity. The groups are sorted by the maximum signal of their terms, with each term within a group subsequently sorted by its individual signal. This grouping is visually highlighted on the right side of the plot.

STRING clustering and gene set descriptions

STRING offers network clustering options for user-submitted gene lists. This feature visually connects nodes that are more interconnected to each other than to other nodes, reflecting the functional modularity of the user's gene list. By grouping proteins that closely interact, clustering aids in the discovery of functional modules and facilitates hypothesis generation. Clustering is based on the connectivity between nodes, incorporating edge weights (combined scores) reflecting the confidence level of the interactions. Only visible edges—those corresponding to active channels with confidence higher than the user-specified cutoff—are considered, ensuring that clustering is consistent with the network being shown.

Users have the choice of two different clustering algorithms. '*K*-means clustering' is implemented via the *k*-medoids algorithm [pyclustering package (37)], allowing users to specify the desired number of clusters (*k*). This method forms clusters based on a distance matrix derived from the most probable paths between nodes. In case of multiple disconnected components within the network, STRING aims for a balanced distribution of cluster sizes by iteratively dividing the largest components first. In contrast, 'MCL clustering', implemented with standalone binaries (release 14-137), automatically generates natural divisions based on the network's inherent structure. It utilizes the Markov cluster algorithm (38) to simulate stochastic flow in graphs, identifying groups of highly interconnected nodes as clusters based on their combined STRING scores. The user-controlled inflation parameter influences the granularity of the clustering. This method is particularly effective in capturing the true modular nature of biological networks, allowing for intuitive groupings that correspond to biological functions and relationships.

Researchers often analyze the biological context of the obtained clusters to interpret and draw conclusions from network data. To aid this process, we have developed a novel gene set naming feature that significantly enhances the interpretability of cluster analyses. This feature automatically assigns the best description for each cluster based on STRING's robust statistical enrichment analysis, simplifying the often cumbersome task of manually interpreting cluster data. It ranks the enriched terms by their enrichment signal (see the 'Improved enrichment analysis' section), choosing the primary, secondary and tertiary names among various enrichment categories. This prioritization ensures that the names reflect the most statistically significant and biologically pertinent attributes of each cluster, providing a clear, immediate understanding of its functional characteristics. The enrichment primarily draws from categories such as Gene Ontology Biological Processes, while incorporating a range of other categories for broader annotations. In case the enrichment analysis yields no significant results or if the cluster contains only one gene, STRING assigns canonical gene names as descriptions to en-

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

sure that clusters are easily identifiable. This functionality extends beyond cluster analysis and is applicable to any gene set. As such, it is also available through an API, complementing our suite of other API methods. The new API function, named *geneset_description*, requires only a set of genes as input and automatically performs enrichment analysis. It returns up to three descriptions—primary, secondary and tertiary—based on relevance and availability, identical to the cluster naming in the user interface. These descriptions are filled sequentially, with the 'primary' always representing the most relevant term. One application of this API is in the stringApp Cytoscape plugin, which utilizes it to automatically overlay the assigned descriptions onto each network cluster after clustering is executed.

Network and protein embeddings

To facilitate the use of STRING in machine learning applications, users can now directly download precomputed ProtT5 (39) sequence and cross-species protein network embeddings derived from STRING, for all eukaryotes, and utilize the embeddings on their own labeled datasets. The embeddings encode information from two aspects. Sequence embeddings can capture protein domains and shorter sequence motifs, while network embeddings can complement these with information on cellular organization such as protein complexes and pathways that may not be evident from sequence data alone.

While protein sequence embeddings are inherently comparable across species due to the universal nature of amino acid sequences, the primary obstacle to using network embeddings is ensuring that network embeddings from different species are directly comparable. This challenge arises because protein networks are independent for each species, and traditional network embedding methods are not designed to address cross-species comparability. The lack of comparability between network embeddings prevents the effective transfer of knowledge and findings across different species. To address this, we developed a method to align eukaryotic network embeddings across species using orthologous relationships based on FedCoder (40). This alignment technique creates a unified embedding space where proteins from different eukaryotic species can be directly compared. As a result, it enhances cross-species protein predictions, particularly in tasks such as subcellular localization and function prediction.

In the protein machine learning field, having precomputed sequence embeddings and cross-species compatible network embeddings from STRING enables researchers to use these resources directly, eliminating the need to calculate embeddings themselves. This strategy not only reduces the energy footprint resulting from redundant calculations across research groups, but also enhances reproducibility in computational biology research promoting more sustainable and consistent scientific practices.

Data availability

The STRING database is freely available online at https://string-db.org/.

Acknowledgements

The authors wish to thank Yan P. Yuan (EMBL Heidelberg) and Dandan Xue (University of Copenhagen) for IT support. Thomas Rattei (University of Vienna) is thanked for extensive computational work toward resolving orthology relations. We thank the CSC—IT Center for Science, Finland, for generous computational resources.

Funding

Swiss Institute of Bioinformatics; Novo Nordisk Foundation [NNF14CC0001, NNF20SA0035590]; European Molecular Biology Laboratory (EMBL Heidelberg); HORIZON EUROPE Marie Skłodowska-Curie Actions [101023676 to K.N.]; Academy of Finland [332844 to F.M. and S.P.]. Funding for open access charge: University of Zurich.

Conflict of interest statement

None declared.

References

1. Marsh,J.A. and Teichmann,S.A. (2015) Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.*, **84**, 551–575.
2. Cesnik,A., Schaffer,L.V., Gaur,I., Jain,M., Ideker,T. and Lundberg,E. (2024) Mapping the multiscale proteomic organization of cellular and disease phenotypes. *Annu. Rev. Biomed. Data Sci.*, **7**, 369–389.
3. Milacic,M., Beavers,D., Conley,P., Gong,C., Gillespie,M., Griss,J., Haw,R., Jassal,B., Matthews,L., May,B., *et al.* (2024) The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.*, **52**, D672–D678.
4. Kanehisa,M., Furumichi,M., Sato,Y., Kawashima,M. and Ishiguro-Watanabe,M. (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*, **51**, D587–D592.
5. Oughtred,R., Stark,C., Breitkreutz,B.-J., Rust,J., Boucher,L., Chang,C., Kolas,N., O'Donnell,L., Leung,G., McAdam,R., *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
6. Del Toro,N., Shrivastava,A., Ragueneau,E., Meldal,B., Combe,C., Barrera,E., Perfetto,L., How,K., Ratan,P., Shirodkar,G., *et al.* (2022) The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.*, **50**, D648–D653.
7. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E., *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
8. Meldal,B.H.M., Perfetto,L., Combe,C., Lubiana,T., Ferreira Cavalcante,J.V., Bye-A-Jee,H., Waagmeester,A., Del-Toro,N., Shrivastava,A., Barrera,E., *et al.* (2022) Complex Portal 2022: new curation frontiers. *Nucleic Acids Res.*, **50**, D578–D586.
9. Lo Surdo,P., Iannuccelli,M., Contino,S., Castagnoli,L., Licata,L., Cesareni,G. and Perfetto,L. (2023) SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.*, **51**, D631–D637.
10. Szklarczyk,D., Kirsch,R., Koutrouli,M., Nastou,K., Mehryary,F., Hachilif,R., Gable,A.L., Fang,T., Doncheva,N.T., Pyysalo,S., *et al.* (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.*, **51**, D638–D646.
11. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T., *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
12. Persson,E., Castresana-Aguirre,M., Buzzao,D., Guala,D. and Sonnhammer,E.L.L. (2021) FunCoup 5: functional association

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Högskolan Biblioteket user on 02 October 2026

networks in all domains of life, supporting directed links and tissue-specificity. *J. Mol. Biol.*, **433**, 166835.
13. Kim,C.Y., Baek,S., Cha,J., Yang,S., Kim,E., Marcotte,E.M., Hart,T. and Lee,I. (2022) HumanNet v3: an improved database of human gene networks for disease research. *Nucleic Acids Res.*, **50**, D632–D639.
14. Huang,J.K., Carlin,D.E., Yu,M.K., Zhang,W., Kreisberg,J.F., Tamayo,P. and Ideker,T. (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.*, **6**, 484–495.e5.
15. Wright,S.N., Colton,S., Schaffer,L.V., Pillich,R.T., Churas,C., Pratt,D. and Ideker,T. (2024) State of the interactomes: an evaluation of molecular networks for generating biological insights. bioRxiv doi: https://doi.org/10.1101/2024.04.26.587073, 29 April 2024, preprint: not peer reviewed.
16. Ontology Consortium,G., Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
17. Liberzon,A., Birger,C., Thorvaldsdóttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
18. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
19. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 5890–5895.
20. Guala,D., Ogris,C., Müller,N. and Sonnhammer,E.L.L. (2020) Genome-wide functional association networks: background, data & state-of-the-art resources. *Brief. Bioinform.*, **21**, 1224.
21. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
22. Mering,C.V. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
23. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
24. Franceschini,A., Lin,J., von Mering,C. and Jensen,L.J. (2016) SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, **32**, 1085–1087.
25. Koutrouli,M., Nastou,K., Piera Líndez,P., Bouwmeester,R., Rasmussen,S., Martens,L. and Jensen,L.J. (2024) FAVA: high-quality functional association networks inferred from scRNA-seq and proteomics data. *Bioinformatics*, **40**, btae010.
26. Porras,P., Barrera,E., Bridge,A., Del-Toro,N., Cesareni,G., Duesbury,M., Hermjakob,H., Iannuccelli,M., Jurisica,I., Kotlyar,M., *et al.* (2020) Towards a unified open access dataset of molecular interactions. *Nat. Commun.*, **11**, 6144.
27. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C., *et al.* (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
28. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.-D.J., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.*, **14**, 1107–1118.
29. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
30. Doncheva,N.T., Morris,J.H., Gorodkin,J. and Jensen,L.J. (2019) Cytoscape StringApp: network analysis and visualization of proteomics data. *J. Proteome Res.*, **18**, 623–632.
31. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M., *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
32. Nastou,K., Mehryary,F., Ohta,T., Luoma,J., Pyysalo,S. and Jensen,L.J. (2024) RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. *Database*, **2024**, baae095.
33. Megill,C., Martin,B., Weaver,C., Bell,S., Prins,L., Badajoz,S., McCandless,B., Pisco,A.O., Kinsella,M., Griffin,F., *et al.* (2021) cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv doi: https://doi.org/10.1101/2021.04.05.438318, 6 April 2021, preprint: not peer reviewed.
34. Moreno,P., Fexova,S., George,N., Manning,J.R., Miao,Z., Mohammed,S., Muñoz-Pomer,A., Fullgrabe,A., Bi,Y., Bush,N., *et al.* (2022) Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.*, **50**, D129–D140.
35. Timmons,J.A., Szkop,K.J. and Gallagher,I.J. (2015) Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol.*, **16**, 186.
36. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.
37. Novikov,A. (2019) PyClustering: data mining library. *J. Open Source Softw.*, **4**, 1230.
38. Van Dongen,S. (2008) Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.*, **30**, 121–141.
39. Elnaggar,A., Heinzinger,M., Dallago,C., Rehawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Steinegger,M., *et al.* (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.
40. Baumgartner,M., Dell'Aglio,D., Paulheim,H. and Bernstein,A. (2023) Towards the Web of Embeddings: integrating multiple knowledge graph embedding spaces with FedCoder. *J. Web Semant.*, **75**, 100741.

Received: September 15, 2024. **Revised:** October 18, 2024. **Editorial Decision:** October 21, 2024. **Accepted:** October 29, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/article/53/D1/D730/7903368 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

OXFORD GENETICS

GENETICS, 2023, 224(1), iyad031
https://doi.org/10.1093/genetics/iyad031
Advance Access Publication Date: 3 March 2023
Knowledgebase & Database Resources

The Gene Ontology knowledgebase in 2023

The Gene Ontology Consortium*

GO Central: Suzi A. Aleksander, James Balhoff, Seth Carbon, J. Michael Cherry, Harold J. Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L. Harris, David P. Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J. Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W. Sternberg, Paul D. Thomas, Kimberly Van Auken
CACAO, EcoliWiki: Jolene Ramsey, Deborah A. Siegele
dictyBase: Rex L. Chisholm, Petra Fey
DisProt, Dept. of Biomedical Sciences, University of Padova, Italy: Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto
ECO: Michelle Giglio, Suvarna Nadendla
FlyBase: Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J. Tabone, Jim Thurmond, Pinglei Zhou
Functional Gene Annotation, Institute of Cardiovascular Science, University College London (London, UK): Saadullah H. Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N. Erdol, Matthew C. Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C.C. Saverimuttu, Renzhi Su, Kate E. Thurlow, Ruth C. Lovering
GREEKC: Colin Logie
JaponicusDB (London, UK): Snezhana Oliferenko
MGI: Judith Blake, Karen Christie, Lori Corbani, Mary E. Dolan, Harold J. Drabkin, David P. Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith
PHI-base: Alayne Cuzick, James Seager
Planteome: Laurel Cooper, Justin Elser, Pankaj Jaiswal
Plant Reactome: Parul Gupta, Pankaj Jaiswal, Sushma Naithani
PomBase: Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood
RGD: Jeffrey L. De Pons, Melinda R. Dwinell, G. Thomas Hayman, Mary L. Kaldunski, Anne E. Kwitek, Stanley J. F. Laulederkind, Marek A. Tutaj, Mahima Vedi, Shur-Jen Wang
Reactome: Peter D'Eustachio
Rhea, Swiss-Prot group, SIB Swiss Institute of Bioinformatics (SIB) (Geneva, Switzerland): Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat
SGD, Department of Genetics, Stanford University (Stanford, CA): Suzi A. Aleksander, J. Michael Cherry, Stacia R. Engel, Kalpana Karra, Stuart R. Miyasato, Robert S. Nash, Marek S. Skrzypek, Shuai Weng, Edith D. Wong
The Arabidopsis Information Resource (TAIR), Phoenix Bioinformatics (Newark, CA, USA): Erika Bakker, Tanya Z. Berardini, Leonore Reiser
UniProt, Swiss-Prot group, SIB Swiss Institute of Bioinformatics (SIB) (Geneva, Switzerland): Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram
UniProt, EMBL-EBI, (Hinxton, UK): Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J. Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru
University at Buffalo, Department of Biomedical Informatics (Buffalo, NY, USA): Alexander D. Diehl
WormBase: Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki
Xenbase: Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn
ZFIN: Sridhar Ramachandran, Leyla Ruzicka, Monte Westerfield

*Corresponding author: Paul D Thomas, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA 90032, USA. Email: pdthomas@usc.edu

Received: December 13, 2022. **Accepted:** February 11, 2023
© The Author(s) 2023. Published by Oxford University Press on behalf of The Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Abstract

The Gene Ontology (GO) knowledgebase (http://geneontology.org) is a comprehensive resource concerning the functions of genes and gene products (proteins and noncoding RNAs). GO annotations cover genes from organisms across the tree of life as well as viruses, though most gene function knowledge currently derives from experiments carried out in a relatively small number of model organisms. Here, we provide an updated overview of the GO knowledgebase, as well as the efforts of the broad, international consortium of scientists that develops, maintains, and updates the GO knowledgebase. The GO knowledgebase consists of three components: (1) the GO—a computational knowledge structure describing the functional characteristics of genes; (2) GO annotations—evidence-supported statements asserting that a specific gene product has a particular functional characteristic; and (3) GO Causal Activity Models (GO-CAMs)—mechanistic models of molecular "pathways" (GO biological processes) created by linking multiple GO annotations using defined relations. Each of these components is continually expanded, revised, and updated in response to newly published discoveries and receives extensive QA checks, reviews, and user feedback. For each of these components, we provide a description of the current contents, recent developments to keep the knowledgebase up to date with new discoveries, and guidance on how users can best make use of the data that we provide. We conclude with future directions for the project.

Keywords: Gene Ontology, gene annotation, gene function, knowledgebase, knowledge graphs

Introduction

Genes encode gene products, often proteins but also noncoding RNA molecules (ncRNAs), that perform functions at the molecular, cellular, and organismal levels. The Gene Ontology (GO) knowledgebase provides a comprehensive, structured, computer-accessible representation of gene function, for genes from any cellular organism or virus. The GO knowledgebase has become a critical component of life science research, supporting analysis of large-scale experiments and biological systems (Duck *et al.* 2016). It is designed to make expert knowledge of gene function accessible for bench scientists as well as computational analyses. The basic model underlying GO is the "molecular biology paradigm" (Ashburner *et al.* 2000; Thomas 2017), in which there are three types (aspects) of functional characteristics used to describe gene function:

- Molecular function (MF): the activities performed by a gene product at the molecular level
- Cellular component (CC): the locations, relative to cellular structures, where MFs are performed
- Biological process (BP): a "biological program" comprising molecular activities acting in concert to achieve a particular outcome; this program can be at the cellular level or at the organism level of multicellular organisms.

The GO knowledgebase consists of three components: the GO, GO annotations, and GO Causal Activity models (GO-CAMs) (Fig. 1). The GO (Fig. 1a) structures our current knowledge of the types of functional characteristics that a gene product may possess into a connected graph-based representation. Each ontology term (called "class" in the field of ontologies) represents a functional characteristic that can be attributed to a gene product. Terms can have relationships between them, such as one term being more specific than another term (also called "subclass"), e.g. *DNA-binding transcription factor activity* is a subclass of *transcription regulator activity*. A GO annotation (Fig. 1b) is an association between a specific gene (or gene product) and a GO term and should be interpreted as a statement that the specified gene product possesses the specified functional characteristic represented by the GO term. Each GO annotation includes the evidence upon which it is based. Because each GO annotation covers only a single characteristic of gene function, multiple GO annotations are generally required to completely describe the function of a gene product. GO-CAMs (Fig. 1c) link multiple GO annotations together to create models of BPs by (1) connecting the activities of more than one gene product together into causal networks and (2) allowing the specification of the biological context (e.g. cell type and tissue type) in which the activities occur.

The GO knowledgebase is large and dynamic

For applications that use the components of the GO knowledgebase, it is crucial that the ontology and associated annotations represent the current state of knowledge and are not just an archive of all public data. Therefore, all aspects of the GO knowledgebase are dynamic (ontology, annotations, GO-CAMs, links to external ontologies, etc.), and citable, versioned updates are released on a monthly basis. Below, we describe each component of the knowledgebase, focusing on recent changes made to improve the resource during the past two years. Statistics and descriptions given here are based on the GO release 2022–11–03 (http://release.geneontology.org/2022-11-03, doi:10.5281/zenodo.7407024).

Ontology

The ontology component of the GO knowledgebase consists of the terms used to describe functional characteristics of gene products, which are linked together by relations into a labeled directed acyclic graph (like a hierarchy but with multiple parentages allowed). It also includes term definitions, synonyms, and relations to terms from external ontologies. The GO is available in different editions, including (1) the "basic" edition, which includes only core relationship types; (2) the core ontology, including additional relationship types; and (3) the "go-plus" edition which also includes relationships to terms in other ontologies. These editions are explained on the GO downloads page http://geneontology.org/docs/download-ontology/. The ontology contains 43,303 terms (Table 2), linked together by 88,099 relationships in the basic edition. When relationships to external terms are included, there are 121,698 relationships; release statistics can be viewed at http://geneontology.org/stats.

The GO is subject to constant review and revision to most accurately model the current biological knowledge. Revision of the ontology includes the addition or obsoletion of terms and reorganization of the relationship structure. New GO terms are added to represent concepts previously missing from the GO in response to published findings, or when a branch of GO is revised. Terms may be obsoleted when unused or inconsistently used in annotation, when they are redundant with other terms, or during revision of specific branches of the ontology.

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

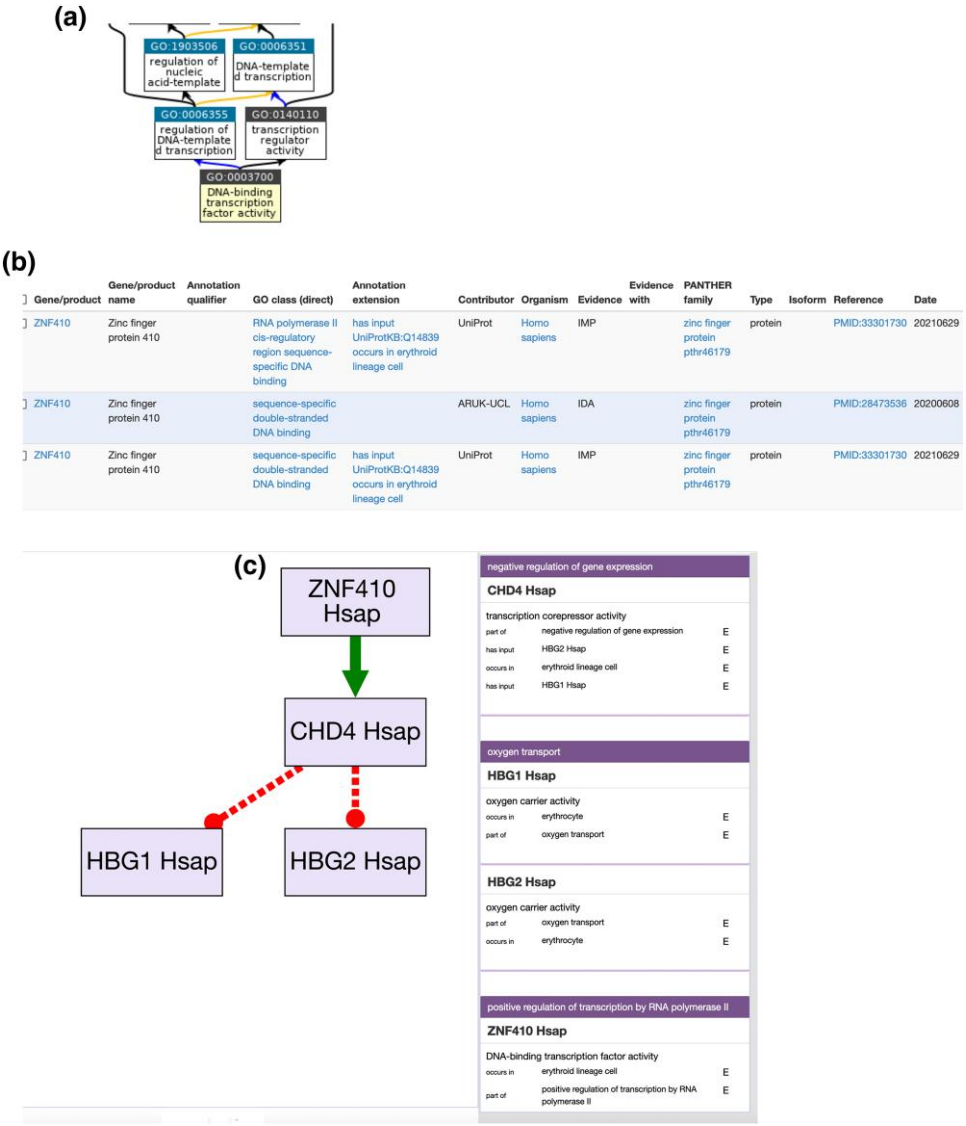


Fig. 1. Examples of the three components of the GO knowledgebase. a) The GO ontology consists of terms, e.g. *DNA binding transcription factor activity*, and relationships between the terms (arrows; black = *is a*, blue = *part of*, and orange = *regulates*). b) GO annotations associate a specific gene product (here, human ZNF410) with GO terms asserting its functional aspects ("GO Class" column, e.g. sequence-specific double-stranded DNA binding) and the evidence for each assertion with its traceable source ("Evidence" and "Reference" columns). c) The GO-CAM model combines individual GO annotations into a model, in this case a very simple model describing how human ZNF410 acts as a transcription factor to positively regulate (denoted by the arrow) transcription of the *CHD4* gene, which in turn acts as a corepressor to repress (denoted by dashed lines) transcription of fetal hemoglobin genes (*HBG1* and *HBG2*) in erythroid lineage cells. In this view, each box in the GO-CAM is labeled with the gene product and species abbreviation for simplicity.

Most of the revisions in the structure of GO are in response to advances in biological knowledge, as well as improvements in the precision of newer experimental approaches. In addition, because many branches of the ontology have grown organically in a bottom-up fashion by accumulating specific individual term requests, we also perform systematic review aimed at improving consistency and clarity while reducing redundancy. Additional revisions are initiated by internal review, and consistency and quality assurance checks. Revisions are also made following feedback from users. Whenever possible, changes are performed in collaboration with expert biocurators or domain specialists; recent examples include blood–brain barrier-related functions (Saverimuttu *et al.* 2021) and transcription factors (Gaudet *et al.* 2021).

At each release, we track all changes and report on our website the number of added, obsoleted, and merged terms in the ontology. Table 1 shows the number of GO terms added and removed (merged or obsoleted) over the past two-year period, for each aspect of GO. In the MF and CC aspects of the ontology, term creation versus term obsoletion have approximately balanced each other, such that the number of terms in these two branches has remained roughly constant. The most significant changes have been in the BP aspect of GO, with a net decrease of over 800 terms.

Many of these revisions result from global reviews of the ontology to address clear inconsistencies in usage and changes in annotation practices. Terms that have been removed from the ontology over the last two years fall into several different categories, including the following:

- **Terms that correspond to phenotypes** and for which the understanding of the process was previously too incomplete to annotate to a different term. Examples include the following: *regulation of spindle density* (GO:0090225) and *age-dependent general metabolic decline* (GO:0007571).

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

- Terms that are combinations of multiple GO terms that **can now be represented more precisely using GO-CAM models**. Examples are *chromatin remodeling in response to cation stress* (GO:0043156) and *regulation of cyclin-dependent protein serine/threonine kinase activity involved in G2/M transition of the mitotic cell cycle* (GO:0031660).
- **Revisions based on updated knowledge**, either by GO editors, by authoritative databases, or in the literature. For example, alpha-taxilin (UniProt:P40222) was originally thought to be the high-molecular weight interleukin-14 (Ambrus *et al.* 1993); an erratum was later published (Ambrus *et al.* 1996) indicating that the open reading frame had been incorrectly predicted. Hence, all terms mentioning interleukin-14 have been obsoleted [*interleukin-14 binding* (GO:0019974), *interleukin-14 production* (GO:0032617), and 5 other terms]. GO terms created from articles that have been retracted and for which no other supporting evidence exists are obsoleted, for example *CDP-acylglycerol O-arachidonoyltransferase activity* (GO:0047193) (Thompson and Zuk 1983). Some terms have been obsoleted from authoritative databases, for example EC:1.3.1.59 was removed from the Enzyme Commission

Table 1. Changes to GO terms in the past two-year period. The ontology has undergone substantial revision and improvement, with nearly 2,000 terms added or removed.

GO aspect	Total number of terms	Added terms	Obsoleted terms	Merged terms[a]
Molecular function	11,271	315	65	143
Cellular component	4,039	34	19	162
Biological process	27,993	217	782	254

[a] Also includes obsoleted terms that have been replaced by another term.

Table 2. Groups contributing literature-based annotations. Includes all annotations traceable to the literature (EXP, including HTP, TAS, NAS, and IC; see http://geneontology.org/docs/guide-go-evidence-codes; see below for information). Direct annotations to the term "protein binding" are listed separately, since without information about interacting partner(s), protein binding represents an activity that most proteins possess, and therefore the GO class itself provides little information (see text for further description). The statistics for groups that have contributed more than 700 manual annotations. Other contributing groups include the following: HGNC, JaponicusDB, PHI-base, PAMGO, JCVI, MENGO, and GDB. Current GO Consortium members are labeled with an asterisk. See http://geneontology.org/docs/annotation-contributors/ for more details.

Group	Organism or area of focus	Number of literature-based annotations, excluding direct protein binding	Number of literature-based annotations directly to protein binding
UniProt* (UniProt: the universal protein knowledgebase 2017)	Human and also a wide variety of organisms not covered by other GOC members	185,121	30,927
MGI* (Bult *et al.* 2019)	Mouse	106,435	8,051
Reactome* (Fabregat *et al.* 2018)	Human pathways	92,178	6
TAIR* (Lamesch *et al.* 2012)	*A. thaliana* (model plant)	64,633	4,695
FlyBase* (Sian *et al.* 2022)	*D. melanogaster* (fruit fly)	55,203	892
UCL*	Human	54,595	2,935
RGD* (Smith *et al.* 2020)	Rat	47,694	1,894
SGD* (Lang *et al.* 2018)	*Saccharomyces cerevisiae* (Baker's yeast)	48,811	165
ZFIN* (Howe *et al.* 2021)	Zebrafish	28,261	488
PomBase* (Harris et al. 2022)	*Schizosaccharomyces pombe* (fission yeast)	26,128	2,201
GeneDB	Microbial pathogens	23,884	756
ComplexPortal* (Meldal *et al.* 2019)	Protein complexes	18,343	0
WormBase* (Davis *et al.* 2022)	*C. elegans* (nematode)	17,171	560
CGD* (Skrzypek *et al.* 2017)	*Candida albicans* (yeast pathogen)	17,113	0
EcoCyc* (Keseler *et al.* 2017)	*E. coli* (bacterium)	13,372	829
AgBase	Agricultural animals, primarily chicken	11,198	1,110
dictyBase* (Basu *et al.* 2015)	*Dictyostelium discoideum* (slime mold)	9,615	844
HPA	Human protein subcellular localization	9,963	0
SynGO (Koopmans *et al.* 2019)	Neuron–neuron synapses	9,552	0
PINC	Human and mouse	6,746	0
MTBBASE	*Mycobacterium tuberculosis* (bacterial pathogen)	6,160	463
IntAct* (Del Toro et al. 2022)	Protein–protein interactions	4,849	216,488
CAFA (Radivojac *et al.* 2013)	Various	4,818	371
CACAO* (Ramsey *et al.* 2021)	Various	4,382	0
AspGD (Cerqueira *et al.* 2014)	*Aspergillus niger* (fungal pathogen)	4,099	0
PseudoCAP (Winsor *et al.* 2005)	*Pseudomonas aeruginosa* (bacterium)	2,323	0
EcoliWiki* (McIntosh *et al.* 2012)	*E. coli* (bacterium)	2,123	55
TIGR	Bacteria	2,150	0
GO_Central*	Various	3,643	160
CollecTF	Bacterial transcription factors	1,850	0
NTNU_SB	Human, mouse, and rat transcription factors	1,733	0
GR	Rice	1,260	0
SGN	Tomato	1,255	0
DisProt	Disordered proteins	933	156
Xenbase* (Fortriede *et al.* 2020)	Xenopus (frog)	731	0

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

database, and the corresponding GO term *1,6-dihydroxy-5-methylcyclohexa-2,4-dienecarboxylate dehydrogenase activity* (GO:0018512) was obsoleted in GO.

- **Single step reactions in the BP aspect of the ontology:** there were many instances in the GO where a MF could be represented as both a MF and a BP, for example "histone kinase activity" and "histone phosphorylation." This was useful when fewer activities were characterized at the molecular level, and the best level of resolution for many experiments was that the gene has some uncharacterized role that led to histone phosphorylation, for example. However, with increasingly detailed molecular data, the redundancy between MF and BP annotations became unnecessary and the value of having a similar term in both aspects of the ontology led to inconsistency. This is an ongoing project and many BP terms still need to be obsoleted for this reason.
- **Terms that refer to more than one ontological aspect**: *ubiquinone biosynthetic process monooxygenase activity* (GO:0015997) included a BP within a MF; *MAP kinase phosphatase activity involved in regulation of innate immune response* (GO:0038078) included a MF within a BP; and *histone deacetylation at centromere* (GO:0031059) represented all three aspects: a MF in the BP branch of the ontology (histone acetylation) that also included CC information (centromere).
- **Misclassified terms**, for example *urea homeostasis* (GO:0097274) and *creatinine homeostasis* (GO:0097273): while these compounds are important medical biomarkers, the normal process that they measure is proper renal function; therefore, these terms have been obsoleted. Annotations have been rehoused under *renal tubular secretion* (GO:0097254) (or one of its children) or removed if the paper supporting the annotation did not allow one to infer the process that affected the circulating levels of urea or creatinine.
- **Reaction mechanisms:** *primary charge separation* (GO:0009766) and *enzyme active site formation* (GO:0018307) were obsoleted because they represent substeps of reactions which are beyond the scope of GO.
- **Protein-modifying activity terms that mention specific substrates**, for example *[cytochrome c]-arginine N-methyltransferase activity* (GO:0016275), which is captured by the more general *arginine N-methyltransferase activity* (GO:0016274). Substrates can be captured with the "has input" relationship in GO-CAM models and in annotation extensions. The exception to this is the histone code: for GO to represent this important mechanism of gene expression and chromatin structure mechanism, specific activities are created for known histone modifications, for example *histone H2AR3 methyltransferase activity* (GO:0070612) and *histone H3T3 kinase activity* (GO:0072354).
- **Experimental assays and nonphysiological substrates**: some experiments are easier to perform using analogs of physiological substrates. Because GO terms should represent *in vivo* functions, we have removed some terms that represent an experiment rather than its biological conclusion. An example is *rubidium ion transport* (GO:0035826): rubidium is used as a tracer for potassium ions (Gill *et al.* 2004) but has no physiological role in itself. Another example is *regulation of nucleosome density* (GO:0060303), which measures the degree of compaction of chromatin, and is a readout for heterochromatin assembly or disassembly.

Concomitantly with these term obsoletions, many new terms have been added to the ontology in the past two years. An example is *molecular condensate scaffold activity* (GO:0140693) for proteins that nucleate condensates that mediate liquid phase transition. This latter term represents a recent advance in the understanding of the organization of cellular biochemistry (Banani *et al.* 2017).

We have also clarified the level of specificity at which MF terms should be represented in GO. For example, we now strive to create GO terms that represent the range of *in vivo* substrate specificity of an enzyme or transporter. This is in contrast to earlier guidelines, in which a GO term was created for each separate molecular substrate tested in a single, isolated experimental assay or result, which could include nonphysiological substrates. With recent improvements in experimental technologies and practices, it is now often possible to annotate with a concept that more closely matches the biological substrate specificity range of a protein. Therefore, while GO makes cross-references to Enzyme Commission (EC) (McDonald and Tipton 2014), Rhea (Bansal *et al.* 2022), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.* 2022), and MetaCyc (Altman *et al.* 2013), GO does not necessarily create a different term for each of the reactions represented in these resources for each substrate on which a MF acts. For example, the GO term *3-oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity* (GO:0004316) represents the fact that the same gene product has a broad specificity toward 3-oxo-acyl groups, and therefore, we have obsoleted the more specific GO terms that refer to only one specific substrate, such as *3-oxo-cis-Delta9-hexadecenoyl-[acp] reductase activity* (GO:0102072), *3-oxo-glutaryl-[acp] methyl ester reductase activity* (GO:0102131), and *3-oxo-pimeloyl-[acp] methyl ester reductase activity* (GO:0102132). For broad specificity enzymes and transporters, the activity on a specific substrate in a specific pathway can be captured by biocurators in a GO-CAM (Thomas *et al.* 2019) or an annotation extension (Gene Ontology Consortium 2010) rather than in a GO term.

Annotations

A GO annotation is a statement asserting that a particular gene or gene product has a particular functional characteristic (GO term); examples are shown in Fig. 1b. New annotations are continually added to the knowledgebase. In the past two years, experimentally supported gene function annotations have been added from over 10,000 scientific papers. As of November 2022, the GO knowledgebase contains experimental knowledge from almost 173,000 papers. GO annotations derived from experimental data are added primarily by the annotation groups in the GO Consortium, which typically curate biological knowledge by organism (Table 2).

GO annotations are also regularly reviewed and may be edited or removed from the knowledgebase for various reasons, particularly when ontology terms are revised (see "Ontology" section above) or when annotations are invalidated by later experimental data. Annotations to terms that will be obsoleted are manually reviewed and annotations are made to a different term whenever possible. For example, when we edited the ontology for histone modifications, over 2,000 annotations to the obsoleted terms were manually reviewed, and histone modifying enzymes were reannotated to the appropriate MF term, while annotations from indirect effects were either removed or reannotated to different, appropriate GO terms. More minor annotation reviews occur regularly.

The Phylogenetic Annotation with the GO project (see below) involves an integrated biocurator review of annotations that has provided additional quality control. The GO user community also plays an important role in identifying incorrect annotations. Because each annotation can be traced to the published paper containing the underlying evidence or describing a method used

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

to infer the annotation, users can quickly verify the accuracy of a given annotation. Potential errors can be reported by clicking on the "Help" link at the top of the GO homepage (http://geneontology.org). In addition, authors of a paper used to create GO annotations can easily retrieve and review all annotations from a given paper and suggest changes; this can be done from the PubMed abstract page (e.g. the PubMed page https://pubmed.ncbi.nlm.nih.gov/20516198/ (Lydeard *et al.* 2010)) by clicking on LinkOut and then the "Gene Ontology" link.

Phylogenetic annotations as a source of highly reviewed annotations

The Phylogenetic Annotation using GO (PAN-GO) project creates a set of biocurator-reviewed, selected GO annotations. The PAN-GO process is described in detail in Gaudet *et al.* 2011. Briefly, using the PAINT software tool, a biocurator reviews all experimentally supported GO annotations collected for all members of a protein family, in the context of a phylogenetic tree from the PANTHER resource (Thomas *et al.* 2022). They then select the most informative and nonredundant GO terms that represent the gene's functional characteristics. Biocurators then model the evolution of these characteristics in the tree by specifying branches along which the GO terms were gained or lost, taking into account events such as duplications, mutations, horizontal gene transfers, and taxonomic specificity. This allows for different members of the same family to be annotated with different GO terms when justified by the experimental data. All PAN-GO annotations can be traced to experimental evidence in one or more related genes. To date, a total of 8,196 protein families (out of 11,719 families with experimental data) have been curated. The PAN-GO curation effort has prioritized human gene-containing families, though many other families have also been curated. As a result, annotation coverage of a genome generally depends on how closely related it is to humans. PAN-GO annotations are available for 82% of human genes (compared with 68% with experimental evidence alone). Other vertebrate genomes have similarly high coverage, with genomes from other taxa covered at lower but still substantial levels (Tables 3 and 4). PAN-GO annotations are updated at each GO release and are included in the standard, downloadable GO annotation files. These annotations can be identified by the "IBA" (inferred from biological ancestor) evidence code and are available for the 142 organisms included in PANTHER gene families (http://pantherdb.org/panther/summaryStats.jsp).

Protein binding and protein-containing complex annotations

We suggest that users should be particularly cautious when using GO annotations directly to the term *protein binding* (GO:0005515; see Table 2). These are highly specific annotations that include the protein binding partner in another field of the annotation (not in the GO term itself) and should not be used in applications such as gene set enrichment analysis. Instead, they are recommended for applications such as protein–protein interaction network construction for human proteins (which represent the vast majority of direct protein binding annotations in the knowledgebase). Since all protein functions encompass some type of binding (to a substrate or to another protein), GO strives to describe the molecular activity of proteins using at least one term that is not only under the *binding* branch of GO; see also the "noncatalytic MF" section above. Therefore, *binding* (GO:0005488) in isolation can be considered a limited functional description and is represented as a distinct branch of GO MF.

Table 3. Genome coverage of PAN-GO annotations. Percentage of protein-coding genes with at least one PAN-GO-reviewed annotation, for different taxonomic groups.

Taxonomic group	Number of individually reviewed, annotated genomes	Gene coverage of annotations
Vertebrates	19	66%–83%
Invertebrates	15	40%–68%
Fungi	14	32%–76%
Plants	40	28%–51%
Protists, alveolates, and amoebae	11	19%–46%
Archaea	8	23%–34%
Bacteria	35	20%–57%

Annotation evidence

All annotations are supported by evidence, comprising two fields in the annotation file (Fig. 1b): an *evidence code* that describes the type of evidence, and a *reference* that lists a persistent identifier for tracing the source (provenance) of the original data. It has often been asserted that the most reliable annotations are those made using an experimental evidence code. However, we suggest that users take into account the type of experimental evidence and the level of review of the annotation (Table 5). Some types of experimental evidence, such as inference from a gene expression pattern (IEP), mutant phenotype (IMP), or genetic interaction (IGI), can often be suggestive of function but not definitive when considered in isolation; other annotations for the same gene are often useful to help interpret these annotations. "High-throughput" evidence codes should be treated with particular care. These codes (beginning with the letter H) denote experiments in which many genes are analyzed at the same time, and these annotations are not individually reviewed by either the paper's authors or GO Consortium biocurator (Attrill *et al.* 2019). Conversely, many nonexperimental evidence types are carefully reviewed by experts. Phylogenetic annotations (IBA evidence code) are based on integration and expert assessment of experimental annotations and thus are *individually reviewed* twice: once in making the annotation from published experimental results and once in the context of all annotations for related genes (Gaudet *et al.* 2011). While annotations using the Inferred from Electronic Annotation (IEA) evidence code are considered automated, most implement expert review of a subset of annotations to minimize false positives (for example, UniRule (MacDougall *et al.* 2021) and InterPro2GO (Paysan-Lafosse *et al.* 2022)). The GOC considers these annotations to be accurate though they are often less specific than other annotations.

GO evidence codes correspond to a subset of the terms found in the Evidence and Conclusion Ontology (ECO) (Nadendla *et al.* 2022). Combinations of particular GO internal references (GO_REFs) and evidence codes are also mapped to specific ECO terms (https://github.com/evidenceontology/evidenceontology/blob/master/gaf-eco-mapping.txt). Users needing to map granular ECO terms to GO evidence code abbreviations can use the mapping file provided by ECO (https://github.com/evidenceontology/evidenceontology/blob/master/gaf-eco-mapping-derived.txt).

GO causal activity models

GO-CAMs are models of causal influences between gene products (Thomas *et al.* 2019) or pathways. More precisely, a GO-CAM links the activities (GO MFs) of gene products together by causal relations that specify the effect of one activity on the other. Each element of a GO-CAM is an instance of an ontology class or other

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Table 4. Number of PAN-GO annotations for selected genomes.

Genome	Total IBA annotations	MF annotations	BP annotations	CC annotations
Danio rerio	82,855	23,049	33,888	25,918
Mus musculus	72,554	19,832	29,947	22,775
Rattus norvegicus	71,276	19,792	28,738	22,746
Homo sapiens	68,695	18,537	28,177	21,981
Gallus gallus	59,293	15,847	24,010	19,436
Xenopus tropicalis	43,232	12,687	16,735	13,810
Arabidopsis thaliana	37,509	12,106	12,922	12,481
Caenorhabditis elegans	31,093	9,021	11,728	10,344
Drosophila melanogaster	30,331	8,628	11,202	10,501
Dictyostelium discoideum	18,966	5,556	6,815	6,595
Saccharomyces cerevisiae	15,675	4,286	5,763	5,626
Schizosaccharomyces pombe	13,723	3,719	4,953	5,051
Escherichia coli	5,681	2,171	1,916	1,594

standard database identifiers, so GO-CAMs are highly structured and amenable to computational analysis. The basic unit of a GO-CAM is a "gene product activity unit," which combines a GO MF annotation (molecular activity), together with GO CC (location) and GO BP (larger functional module) annotations that provide the biological context of the activity. The context can be further specified with other ontologies to capture the cell type [using the Cell Type Ontology (Diehl *et al.* 2016)], tissue/anatomical location (using several different ontologies depending on the species, e.g. Uberon (Mungall *et al.* 2012) for most vertebrates, other metazoan ontologies such as the *Drosophila* anatomy ontology (Costa *et al.* 2013), *Caenorhabditis elegans* anatomy ontology (Lee and Sternberg 2003), or nonanimal ontologies as the Plant Ontology (Cooper and Jaiswal 2016), or a temporal period (e.g. GO biological phase). Activity units are linked together by causal relationships from the Relations Ontology (Smith *et al.* 2005) to capture how they interact to impact larger pathways, modules, or processes.

As of November 2022, GO Consortium annotation groups have created over 300 GO-CAM models that describe molecular pathways (defined as containing at least three distinct gene product activities linked into a causal chain). These models reflect curation priorities of the contributing groups. Most of the available GO-CAMs are for processes in human or mouse, with a limited number in zebrafish, *Drosophila melanogaster*, and *C. elegans*. Many of the human GO-CAMs describe chromatin-mediated regulation of gene expression and immune response pathways, while the mouse GO-CAMs focus on metabolic and signaling pathways. GO-CAMs are accessible from the GO website homepage, by clicking on the "Browse GO-CAMs" link. GO-CAMs can be viewed as pathway diagrams (Fig. 2) and are currently available on GitHub at https://github.com/geneontology/noctua-models.

Community collaborations

The GO Consortium collaborates with experts in specific areas of molecular and cellular biology to systematically update and improve their representation in the ontology and the corresponding GO annotations and GO-CAMs. We recently revised the representation for transcription factors and transcriptional regulation in collaboration with the GREEKC Consortium (Kuiper *et al.* 2022). Additional collaborative projects include working with the DisProt project (Quaglia *et al.* 2022) on improving the ontology and annotations for intrinsically disordered proteins (IDPs); revising processes that involve molecular pathways between interacting species, such as viral infection processes; and integrating the GO and annotations with external biochemical databases.

In 2021, the GO started a collaboration with DisProt (https://disprot.org/)—the gold standard database of manually curated annotations from the literature for IDPs. IDPs lack a stable three-dimensional structure and are characterized by highly flexible and unstructured segments, i.e. intrinsically disordered regions (IDRs). DisProt has developed a custom ontology, the Intrinsically Disordered Proteins Ontology (IDPO), and used it to annotate the structural states of IDPs. The GO Consortium and DisProt have collaborated to refactor IDPO and map the IDPO terms to GO terms whenever possible (those related to functions and interactions of IDPs). The collaboration between the GO Consortium and the DisProt database included the creation and addition of new GO terms to align with already existing IDPO terms that were not yet available in GO. These newly created terms also include the *MF activator* (GO:0140677) and *MF inhibitor* (GO:0140678) terms, used to annotate MF regulators that activate/inhibit or increase/decrease the activities of their targets via noncovalent binding that does not result in covalent modification to the target. This collaboration resulted in more accurate and detailed annotation of the modes of action of IDPs, e.g. *localization* (GO:0051179, IDPO:00010) and *DNA binding* (GO:0003677, IDPO:00065), as well as providing GO annotations. Currently, more than 1,000 expert-curated annotations from DisProt are available in the GO knowledgebase, comprising more than 860 MFs, 200 BPs, and 10 CC annotations. The only terms in IDPO that could not be mapped to GO were those describing self-regulatory (e.g. *self-activation* and *self-inhibition*) and intrinsic disorder-specific functions (i.e. *entropic chains*), so these annotations are available only in DisProt.

Multiorganism interactions

A group that includes experts from within and outside the GO Consortium has been working together to improve and simplify the representation of interactions between organisms, including medically and agriculturally important host–pathogen interactions. Examples of these interactions include how a symbiont such as a virus enters its host, how the host's immune response recognizes and defends the body against a potentially harmful organism, and also beneficial interactions such as how plants form a symbiosis with nitrogen-fixing bacteria. The goal of this project is to revise the host–symbiont branch of GO *BP* to reflect the current scientific knowledge in the field and to ensure that genes are properly annotated to the new ontology terms and structure, building on previous work undertaken as part of the PAMGO consortium (Tyler *et al.* 2009). Symbionts in GO are broadly defined to include pathogens that infect a host organism. We expect that this revision will improve GO-based analyses of molecular studies of pathogens, the mechanisms by which they infect host cells, and host response processes. A major change is that the branch of GO under *BP involved in interspecies interaction between organisms* (GO:0044419) has been reorganized. It now reflects important

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Table 5. GO evidence codes + reference combinations. Users should consider both the type of evidence and the review level. GO internal references (starting with GO_REF:) describe specific annotation methods and are available at https://github.com/geneontology/go-site/tree/master/metadata/gorefs/README.md.

Evidence code	Reference	Evidence-type description	Review-level description
Inferred from direct assay (IDA)	Scientific publication providing experimental data	Experimental, most direct evidence for function	Individually expert-reviewed
Inferred from mutant phenotype (IMP)	Scientific publication providing experimental data	Experimental, from a perturbation in the normal function	Individually expert-reviewed
Inferred from genetic interaction (IGI)	Scientific publication providing experimental data	Experimental, from perturbations in normal functions of more than one gene	Individually expert-reviewed
Inferred from expression pattern (IEP)	Scientific publication providing experimental data	Experimental, used only for biological process annotations, from comparison with genes of known function	Individually expert-reviewed
Inferred from protein interaction (IPI)	Scientific publication providing experimental data	Experimental, used only for annotations to protein binding terms	Individually expert-reviewed
Inferred from high throughput direct assay (HDA)	Scientific publication providing experimental data	Experimental (high throughput), direct	Not individually reviewed; expert review methodology to exclude high false positive rate observations
Inferred from high throughput mutant phenotype (HMP)	Scientific publication providing experimental data	Experimental (high throughput), mutant phenotype	Not individually reviewed; expert review to exclude high false positive rate observations
Inferred from high throughput genetic interaction (HGI)	Scientific publication providing experimental data	Experimental (high throughput), genetic interaction	Not individually reviewed; expert review to exclude high false positive rate observations
Inferred from high throughput expression pattern (HEP)	Scientific publication providing experimental data	Experimental (high throughput), expression pattern	Not individually reviewed; expert review to exclude high false positive rate observations
Inferred from biological ancestor (IBA)	(Gaudet *et al.* 2011)	Homology, from experimental evidence propagated through a phylogenetic tree and/or from direct experimental evidence	Individually expert-reviewed in the context of all experimental annotations for related genes
Inferred from sequence similarity (ISS)	Scientific publication providing sequence similarity evidence	Homology, from experimental evidence propagated from one gene to one related gene, asserted in the publication	Individually expert-reviewed in the context of all experimental annotations for related genes
Inferred from sequence similarity (ISS)	GO_REF:0000024	Homology, from experimental evidence propagated from one gene to one related gene, asserted by a biocurator	Individually expert-reviewed
Inferred from sequence orthology (ISO)	Scientific publication providing orthology evidence	Homology, from experimental evidence propagated from one gene to one orthologous gene	Individually expert-reviewed
Inferred from sequence Orthology	GO_REF:0000008	Homology, from experimental evidence propagated from one mammalian gene to one orthologous mouse gene	Individually expert-reviewed
ISO, Inferred from Sequence orthology (ISO)	GO_REF:0000024	Homology, from experimental evidence propagated from one gene to one orthologous gene	Individually expert-reviewed
Inferred from sequence orthology (ISO)	GO_REF:0000096	Homology, from experimental evidence propagated from one gene to one orthologous gene among human, mouse, and rat orthologs	Not individually reviewed; orthology manually reviewed
Inferred from sequence orthology (ISO)	GO_REF:0000101	Homology, from experimental evidence propagated from one gene to one orthologous gene	Not individually reviewed; orthology computed using OrthoMCL
Inferred from electronic annotation (IEA)	GO_REF:0000107	Homology, from experimental evidence propagated from one gene to one orthologous gene	Not individually reviewed; one-to-one orthology computed using Ensembl Compara phylogenetic trees
Inferred from electronic annotation (IEA)	GO_REF:0000002	Homology, from a hit to an InterPro signature	Not individually reviewed; expert review of annotations of signatures to ensure low or no false positives
Inferred from electronic annotation (IEA)	GO_REF:0000003	Imported from another resource, from mapping an EC number assigned in UniProt	Not individually reviewed; expert review of mappings; EC assignments are manually reviewed for Swiss-Prot and computationally inferred for TrEMBL
Inferred from electronic annotation (IEA)	GO_REF:0000004	Imported from another resource, from mapping a manually assigned Swiss-Prot keyword	Not individually reviewed; expert review of keywords and mappings
Inferred from	GO_REF:0000104		Not individually reviewed; expert curation

(continued)

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Table 5. (continued)

Evidence code	Reference	Evidence-type description	Review-level description
electronic annotation (IEA)		Homology, from manually curated UniRule	of UniRules to ensure low or no false positives
Inferred from electronic annotation (IEA)	GO_REF:0000108	Logical assertion using the ontology, from asserted relation between different aspects of GO	Not individually reviewed; expert curation of ontology links
Inferred from electronic annotation (IEA)	GO_REF:0000117	Computational, from machine learning	Not individually reviewed; assigned by machine learning from curated training sets
Traceable author statement (TAS)	Scientific publication citing original data	From a published statement referencing experimental evidence in a different paper	Individually reviewed
Nontraceable author statement (NAS)	Scientific publication with general biological knowledge statement	From an unreferenced published statement	Individually reviewed

concepts such as the types of biological programs used by symbionts to enable infection and by hosts to prevent or manage infection, such as *disruption of CC of another organism* (GO:0140975), *formation of structure involved in a symbiotic process* (GO:0044111), *killing of cells of another organism* (GO:0031640), and *modulation of process of another organism* (GO:0035821). Each of these terms has multiple, more specific subclass terms.

One challenge in this area was that some previous GO annotations for pathogen genes used terms that apply to normal host processes, such as regulation of defense response processes. Thus, it was not clear whether the pathogen gene was regulating its own defense process or that of a host. With the new ontology terms and structure, these distinctions are clear for both GO biocurators and users of GO. In general, it was important to clearly represent that certain symbiont-initiated processes hijack various host cellular processes. This includes mechanisms to enter and exit the cell, either by binding to host membrane proteins or using the intracellular transport machinery and using the host cellular machinery for genome replication, as well as transcription and translation. We have obsoleted terms that do not clearly distinguish hijacking with the functions that a host gene performs for the host organism, such as *dissemination or transmission of symbiont from host by vector* (GO:0044008) and *positive regulation of viral release from host cell* (GO:1902188). Conversely, a pathogenic symbiont triggers innate responses in the host that are not the evolved role of these symbiont proteins, such as *induction by symbiont of host cytokine production* (GO:0036523) and *pathogen-associated molecular pattern-dependent induction by symbiont of host innate immune response* (GO:0052033)—these are not functions that a symbiont protein performs to enable its own survival and reproduction.

Integration with biochemical knowledgebases

For accurate representation of biochemical aspects of gene function, we work closely with the Rhea database of reactions (Bansal *et al.* 2022) and the ChEBI ontology of chemical entities (Hastings *et al.* 2016). Rhea provides precise representations of *in vivo* biochemical reactions, including precise chemical entity participants and their stoichiometry. Rhea uses ChEBI terms to represent chemical entities in a standardized, consistent manner. The Rhea database overlaps in content with the catalytic activity branch of the GO but provides additional detailed reaction information and in some cases provides additional specificity. We have improved GO mappings to Rhea, which now covers 4,399 GO catalytic activities (in the MF branch of GO). These mappings allow for nonexact matches when the chemical specificity differs between GO and Rhea. For example, Rhea has two reactions, each referring to a different type of beta glucoside (RHEA:69647 and RHEA:69655, narrow match), whereas GO:0008422, beta-glucosidase activity, covers both substrates, as no known enzyme is specific for just one of them. We have recently used the Rhea-GO mappings to include additional linkages between GO MF terms and ChEBI terms in the go-plus release (see below). Previously, ChEBI terms were linked only to general terms in the GO BP branch (e.g. between folate transport and folate), but the additional Rhea linkages have added a total of 4,334 distinct chemical entities linked via 20,307 relationships. The extensive linkage to chemical entities opens opportunities for using GO in other applications, e.g. metabolomics analyses.

Accessing and downloading GO data

Browsing GO and its annotations

GO and associated annotations can be searched directly from the GO home page (http://geneontology.org/), queried using the AmiGO browser (http://amigo.geneontology.org/amigo) or the QuickGO tool (https://www.ebi.ac.uk/QuickGO/) (Munoz-Torres and Carbon 2017). Gene set enrichment analysis is also directly accessible from the GO home page, which launches the PANTHER gene analysis tool at http://pantherdb.org/webservices/go/overrep.jsp (Mi *et al.* 2019).

Ontology downloads

GO provides three editions of the ontology on the download page (http://geneontology.org/docs/download-ontology/) to accommodate various applications: go-basic, go, and go-plus (Table 6). All GO terms, including obsolete terms and term metadata such as definitions, cross-references, and synonyms, are available in all three editions. These editions differ in the set of relations they contain:

- **go-basic** contains the types of information that has been available for GO from the beginning of the project; hence, it only contains is a, part of, regulates, negatively regulates, and positively regulates relationships and excludes relationships that cross different aspects (BP, MF, or CC) of the ontology. This edition of the ontology is guaranteed to be acyclic and can safely be used to selectively propagate annotations across any relation. It is recommended for most GO-based software tools.
- **go** additionally includes has part and occurs in relationships that link terms across different aspects of the ontology (for example, a BP can have a *has part* relation to a MF term or

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

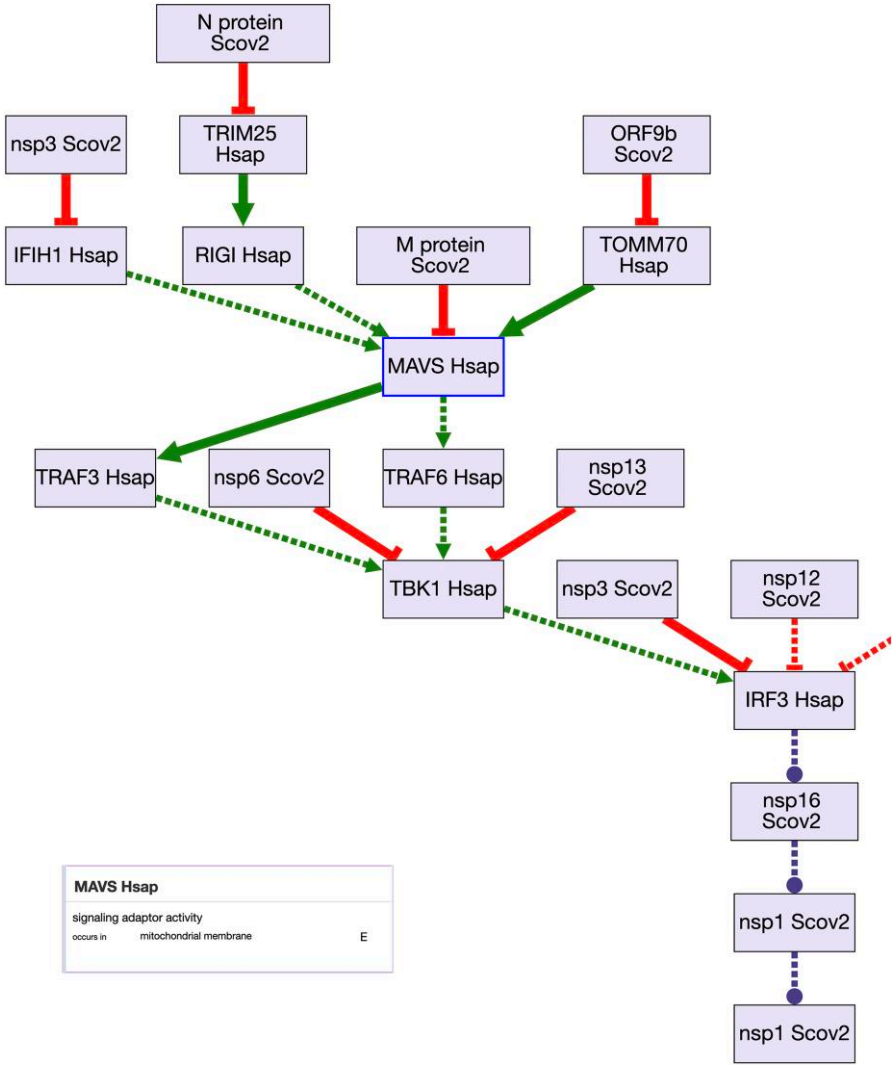


Fig. 2. GO-CAM model of the SARS-CoV2—host interactions as displayed using the GO-CAM Pathway Widget (code available at https://github.com/geneontology/wc-gocam-viz) on the Alliance of Genome Resources gene pages (https://www.alliancegenome.org/gene/HGNC:20144#pathways). The model includes proteins from both humans (Hsap) and the SARS-CoV-2 virus (Scov2). A simplified representation of the causal model is shown on the main figure, which is simplified by labeling each activity with the gene and organism. The model includes many additional details, which are displayed as "cards;" the information for MAVS activity (inset) which normally acts as a signaling adaptor located in the mitochondrial membrane. MAVS activity is suppressed directly by the SARS-CoV-2M protein and indirectly by other SARS-CoV-2 proteins. Each of the "E" symbols on the right-hand side can be clicked to see the evidence for each assertion in the model.

an *occurs in* relation to a CC). This edition is not acyclic and annotations should not be propagated across all the relationship types that it contains. This edition should not be used in most software tools that rely on the GO.

- **go-plus** is the fully axiomatized edition of the ontology and includes cross-ontology relationships to external ontologies including ChEBI, Cell Ontology, and Uberon.

Ontology subsets (GO slims)

GO subsets are condensed versions of the GO containing a portion of the terms, which are specified by tags within the ontology files that indicate if a given term is a member of a particular subset. GO subsets are particularly useful for providing a global overview of the functions of all the genes in a genome, and even for all the functions of a single gene. GO subsets are particularly useful for providing a global overview of the range of functions and processes found in a given clade or organism's genome. We have recently revised the "GO Generic subset," a subset maintained by the GO Consortium that aims to be general and applicable to any species. We have tested that the subset covers as many gene products as possible in various organisms (human, *D. melanogaster*, fission yeast, *Arabidopsis thaliana*, and *Escherichia coli*) with as little redundancy as possible. This new GO Generic Subset contains 75 BP terms, 40 MF terms, and 29 CC terms. The GO generic subset can be accessed at http://current.geneontology.org/ontology/subsets/goslim_generic.obo. Versions in .owl, .json, and .tsv are also available from http://current.geneontology.org/ontology/subsets/index.html.

As part of the Alliance of Genome Resources, we have developed a widget that provides a graphical visualization of a gene's function in a "ribbon"-like display (Fig. 3). The widget can be customized to use any GO subset and uses the goslim_agr subset by default. This widget is implemented in the Alliance gene pages and in the UniProt entry pages. It accesses GO annotations using the GO API (application programming interface) and can be easily added to any webpage.

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Table 6. GO editions. Editions are distinguished by the relations and metadata that they include. All editions are updated at each GO release. External ontologies used in GO include the following: ChEBI, Uberon (Haendel *et al.* 2014), Relation Ontology (Smith *et al.* 2005), Cell Ontology (Diehl *et al.* 2016), Sequence Ontology (Mungall, Batchelor and Eilbeck 2011), Dicty Anatomy, CARO (Haendel *et al.* 2008), Fungal Anatomy Ontology (Fungal-Anatomy-Ontology, 2020), Plant Ontology (Walls *et al.* 2019), PATO (Gkoutos, Schofield and Hoehndorf 2018), and Protein Ontology (Natale *et al.* 2017).

GO edition	Format(s)	Relations included	Links to other ontologies
go-basic	OBO	Is a, part of, regulates, negatively regulates, and positively regulates	Not available
go	OBO and OWL-RDF/XML	Same as go-basic, plus has part, and occurs in	Not available
go-plus	OWL-RDF/XML		ChEBI, Uberon, Cell Ontology, Sequence Ontology, Dicty Anatomy, CARO, Fungal Anatomy Ontology, Plant Ontology, PATO, and Protein Ontology

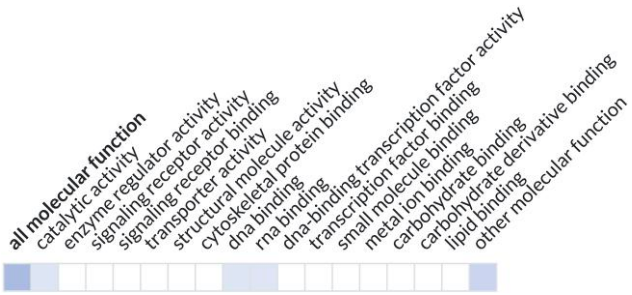


Fig. 3. Alliance ribbon view for the yeast RPB7 gene. High-level GO categories annotated are shown in shaded squares (https://www.alliancegenome.org/gene/SGD:S000002812); darker shading indicates more annotations in that category.

GO annotations

The two major sites for downloading GO annotations are geneontology.org and UniProt-GOA. geneontology.org is the website developed by the GO Consortium. This downloads site (http://current.geneontology.org/products/pages/downloads.html) provides a total of 7.5 million human and model organism annotations contributed by multiple groups. It contains all manually reviewed GO annotations and electronic (computationally predicted) annotations for the most commonly used organisms. For model organisms, all annotations use gene identifiers from the authoritative database (for example, FlyBase (FBgn), WormBase (WBGene), and SGD (S)). Human and other organisms without an authoritative dedicated database are represented by UniProtKB accession numbers. For these organisms, the GO website provides annotations to UniProt reference proteomes (https://www.uniprot.org/help/reference_proteome), which are generally one entry per gene, thus limiting redundancy in annotations. UniProt-GOA (https://www.ebi.ac.uk/GOA/uniprot_release) contains 1 billion annotations for all entries in UniProt (1,264,340 taxa), covering both reviewed (Swiss-Prot) entries of UniProt, and unreviewed (TrEMBL) entries. All annotations for the model organism genes are converted to UniProt protein identifiers. For most organisms, all annotations are electronic annotations generated via various pipelines (see above for evidence codes and references for different methods). In addition to these resources, GO annotations are also viewable in a number of biological databases, including model organism databases, UniProt (UniProt: the universal protein knowledgebase 2017), NCBI (Sayers *et al.* 2020), and The Alliance of Genome Resources (Alliance of Genome Resources Consortium 2022). These sites show GO annotations in the broader context of a gene product's expression pattern, phenotypes, metabolic and signaling pathways, etc.

Conclusions and future directions

The extensive and wide-ranging use of the GO knowledgebase, evidenced by its recent, peer-reviewed designation as a Global Core Biodata Resource (https://globalbiodata.org/scientific-activities/global-core-biodata-resources/), demands its continued development and expansion. We are focusing on several high-priority areas of development for the near future. For pathways, we will continue to accumulate GO-CAM models. The UniProt/Swiss-Prot curation team has ramped up their production of GO-CAM models and we expect to add models at a rapid rate. In parallel, we have started converting Reactome pathways into GO-CAMs (Good *et al.* 2021) and expect to release GO-CAM representations of most Reactome metabolic pathways in the near future. This will provide a complementary, causal flow representation of the chemical reaction-centered representation in Reactome. Conversion of Reactome signaling pathways is more challenging and will be released somewhat later. We are also working on converting the YeastPathways resource (https://pathways.yeastgenome.org) into GO-CAMs, making a large number of yeast metabolic pathways available. The increasing number of GO-CAM models will allow us to expand on the utility of these highly structured pathway and process representations. Some potential areas are automated pathway visualization, using the causal links and more granular gene sets to enhance enrichment analysis, and better generation of automated descriptions of gene function (e.g. Kishore *et al.* 2020).

With respect to ontology development, in addition to continuing to revise the ontology in response to recent discoveries, we see an immediate need for clearly delineating the level of biological organization at which a function is described. This includes distinguishing MFs from BPs, and distinguishing BPs that occur at the level of individual cells, versus those that occur at the level of multicellular organisms. For example, the term "homeostasis"—the maintenance of a roughly steady level of a molecule or ion—is used very broadly in the literature to refer to both processes that maintain a steady-state level within a cell and processes that maintain a steady state in blood or other fluid that is transported within a multicellular organism. Even in some publications, it is difficult to know which type of homeostasis is being tested.

We will continue to make the GO knowledgebase easier to use and more community-driven. One near-term priority is to make annotations available for download by species, with a single identifier for each distinct gene. We are also planning to create quick-

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

start guides for common GO use cases, in both written and video forms. The immense user base of the GO and the need for much improvement and extension drives us to consider how to expand the number of people that contribute to the GO. From its inception, the GO has been a large, open, community project. However, we are planning additional routes through which the broader GO user community can contribute their expert feedback and knowledge to GO, improving the resource for all users. For now, users are encouraged to contact the GO Helpdesk (http://help.geneontology.org/) with any questions or to report any GO ontology terms or annotations that may be inaccurate or difficult to interpret.

Data availability

All GO code and resources are freely available for download and reuse. Software (https://github.com/geneontology/) is under the BSD 3-Clause open-source license. Downloads are available under the CC BY 4.0 license from http://geneontology.org/docs/downloads/

Funding

The core funding for the GOC is from the National Human Genome Research Institute with co-funding from the National Institute of General Medical Sciences (U41HG002273, U24HG012212). Curation activities are supported by National Human Genome Research Institute grants U24HG002659 (ZFIN), U24HG002223 (WormBase), U41HG000739 (FlyBase), U24HG001315 (SGD), U24HG000330 (MGD), U24HG012198 (Reactome curation), and U24HG011851 (Reactome—GO harmonization) and grant R01HL064541 from the National Heart, Lung and Blood Institute (RGD). Additional funding for GO curation at FlyBase is provided by UK Medical Research Council Award MR/W024233/1. PomBase is supported by Wellcome Trust 218236/Z/19/Z. Xenbase is supported by grant P41 HD064556 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Functional Gene Annotation, University College London, is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Planteome and Plant Reactome are supported by National Science Foundation awards #1340112, #1127112, and USDA-ARS. Some software development was funded by National Human Genome Research Institute, grant U24HG010859 (Alliance of Genome Resources Central). The TAIR project is funded by academic, institutional, corporate, and individual subscriptions; TAIR is administered by the 501(c)(3) nonprofit Phoenix Bioinformatics. CJM, SC, and SM were supported in part by Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy Contract No. DE-AC02-05CH11231. UniProt is funded by the National Human Genome Research Institute (NHGRI), Office of Director (OD/DPCPSI/ODSS); National Institute of Allergy and Infectious Diseases (NIAID), National Institute on Aging (NIA), National Institute of General Medical Sciences (NIGMS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Eye Institute (NEI), National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (U24HG007822); Biotechnology and Biological Sciences Research Council (BB/T010541/1, BB/S01781X/1); Open Targets; Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI); and European Molecular Biology Laboratory core funds.

Conflicts of interest

The authors declare no conflicts of interest.

Summary

The Gene Ontology (GO) knowledgebase has broad applications in genetic and genomic research and has been continually updated and improved for more than 20 years. We describe the latest improvements to the GO resource and give an overview of the current contents of the knowledgebase. We also include guidance on the use of GO, annotations, and the growing number of pathways represented as GO-Causal Activity Models (GO-CAMs).

Literature cited

Alliance of Genome Resources Consortium. Harmonizing model organism data in the Alliance of Genome Resources. Genetics. 2022;220:iyac022. doi:10.1093/genetics/iyac022.

Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013;14(1):112. doi:10.1186/1471-2105-14-112.

Ambrus JL Jr, Pippin J, Joseph A, Xu C, Blumenthal D, Tamayo A, Claypool K, McCourt D, Srikiatchatochorn A, Ford RJ. Identification of a cDNA for a human high-molecular-weight B-cell growth factor. Proc Natl Acad Sci U S A. 1993;90(13):6330–6334. doi:10.1073/pnas.90.13.6330.

Ambrus JL Jr, Pippin J, Joseph A, Xu C, Blumenthal D, Tamayo A, Claypool K, McCourt D, Srikiatchatochorn A, Ford R. Identification of a cDNA for a human high molecular-weight B-cell growth factor. Proc Natl Acad Sci U S A. 1996;93(15):8154. doi:10.1073/pnas.93.15.8154-b.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–29. doi:10.1038/75556.

Attrill H, Gaudet P, Huntley RP, Lovering RC, Engel SR, Poux S, Van Auken KM, Georghiou G, Chibucos MC, Berardini TZ, *et al.* Annotation of gene product function from high-throughput studies using the Gene Ontology. Database. 2019;2019:baz007. doi:10.1093/database/baz007.

Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017;18(5):285–298. doi:10.1038/nrm.2017.7.

Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, *et al.* Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res. 2022;50(D1):D693–D700. doi:10.1093/nar/gkab1016.

Basu S, Fey P, Jimenez-Morales D, Dodson RJ, Chisholm RL. Dictybase 2015: expanding data and annotations in a new software environment. Genesis. 2015;53:523–534. doi:10.1002/dvg.22867.

Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Anagnostopoulos A, Asabor R, Baldarelli RM, Beal JS, Bello SM, *et al.* Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2019;47(D1):D801–D806. doi:10.1093/nar/gky1056.

Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, Miyasato SR, Binkley J, Orvis J, Shah P, *et al.* The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. 2014;42(D1):D705–D710. doi:10.1093/nar/gkt1029.

Cooper L, Jaiswal P. The Plant Ontology: A Tool for Plant Genomics. Methods Mol Biol. 2016; 1374:89–114. doi:10.1007/978-1-4939-3167-5_5

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

Costa M, Reeve S, Grumbling G, Osumi-Sutherland D. The Drosophila anatomy ontology. J Biomed Semantics. 2013; 4:32. doi:10.1186/2041-1480-4-32

Davis P, Zarowiecki M, Arnaboldi V, Becerra A, Cain S, Chan J, Chen WJ, Cho J, da Veiga Beltrame E, Diamantakis S, *et al.* Wormbase in 2022-data, processes, and tools for analyzing Caenorhabditis elegans. Genetics. 2022;220(4):iyac003. doi:10.1093/genetics/iyac003.

Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, Perfetto L, How K, Ratan P, Shirodkar G, *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res. 2022;50(D1):D648–D653. doi:10.1093/nar/gkab1006.

Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7(1):44. doi:10.1186/s13326-016-0088-7.

Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, Stevens R. A survey of bioinformatics database and software usage through mining the literature. PLoS One. 2016;11(6):e0157989. doi:10.1371/journal.pone.0157989.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, *et al.* The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–D655. doi:10.1093/nar/gkx1132.

Fortriede JD, Pells TJ, Chu S, Chaturvedi P, Wang DZ, Fisher ME, James-Zorn C, Wang Y, Nenni MJ, Burns KA, *et al.* Xenbase: deep integration of GEO & SRA RNA-seq and ChIP-seq data in a model organism database. Nucleic Acids Res. 2020;48:D776–D782. doi:10.1093/nar/gkz933.

Fungal-Anatomy-Ontology. A Structured Controlled Vocabulary for the Anatomy of Fungi. 2020. https://github.com/obophenotype/fungal-anatomy-ontology.

Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011;12(5):449–462. doi:10.1093/bib/bbr042.

Gaudet P, Logie C, Lovering RC, Kuiper M, Lægreid A, Thomas PD. Gene Ontology representation for transcription factor functions. Biochim Biophys Acta Gene Regul Mech. 2021;1864(11–12):194752. doi:10.1016/j.bbagrm.2021.194752.

Gene Ontology Consortium. The gene ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010;38(suppl_1):D331–D335. doi:10.1093/nar/gkp1018.

Gill S, Gill R, Wicks D, Despotovski S, Liang D. Development of an HTS assay for Na+, K+-ATPase using nonradioactive rubidium ion uptake. Assay Drug Dev Technol. 2004;2(5):535–542. doi:10.1089/adt.2004.2.535.

Gkoutos GV, Schofield PN, Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. Brief Bioinform. 2018;19(5):1008–1021. doi:10.1093/bib/bbx035.

Good BM, Van Auken K, Hill DP, Mi Huaiyu, Carbon Seth, Balhoff JP, Albou LP, Thomas PD, Mungall CJ, Blake JA, *et al.* Reactome and the Gene Ontology: digital convergence of data resources. Bioinformatics. 2021;37(19):3343–3348. doi:10.1093/bioinformatics/btab325.

Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, *et al.* Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J Biomed Semantics. 2014;5(1):21. doi:10.1186/2041-1480-5-21.

Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLV, Mungall CJ, Smith B. CARO—the common anatomy reference ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy Ontologies for Bioinformatics: Principles and Practice. London: Springer London; 2008. p. 327–349.

Harris MA, Rutherford KM, Hayles J, Lock A, Bähler J, Oliver SG, Mata J, Wood V. Fission stories: using PomBase to understand *Schizosaccharomyces pombe* biology. Genetics 2022;220:iyab222. doi:10.1093/genetics/iyab222.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 2016; 44(D1):D1214–9. doi:10.1093/nar/gkv1031.

Howe DG, Ramachandran S, Bradford YM, Fashena D, Toro S, Eagle A, Frazer K, Kalita P, Mani P, Martin R, *et al.* The Zebrafish Information Network: major gene page and home page updates. Nucleic Acids Res. 2021;49(D1):D1058–D1064. doi:10.1093/nar/gkaa1010.

Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG For taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2022;51(D1):D587–D592. doi:10.1093/nar/gkac963.

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, *et al.* The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res. 2017;45(D1):D543–D550. doi:10.1093/nar/gkw1003.

Kishore R, Arnaboldi V, Van Slyke CE, Chan J, Nash RS, Urbano JM, Dolan ME, Engel SR, Shimoyama M, Sternberg PW, *et al.* Automated generation of gene summaries at the Alliance of Genome Resources. Database. 2020;2020:baaa037. doi:10.1093/database/baaa037.

Koopmans F, van Nierop P, Andres-Alonso M, Byrnes A, Cijsouw T, Coba MP, Cornelisse LN, Farrell RJ, Goldschmidt HL, Howrigan DP, *et al.* SynGO: an evidence-based, expert-curated knowledge base for the synapse. Neuron. 2019;103(2):217–34.e4. doi:10.1016/j.neuron.2019.05.002.

Kuiper M, Bonello J, Fernández-Breis JT, Bucher P, Futschik ME, Gaudet P, Kulakovskiy IV, Licata L, Logie C, Lovering RC, *et al.* The gene regulation knowledge commons: the action area of GREEKC. Biochim Biophys Acta Gene Regul Mech. 2022;1865(1):194768. doi:10.1016/j.bbagrm.2021.194768.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(D1):D1202–D1210. doi:10.1093/nar/gkr1090.

Lang OW, Nash RS, Hellerstedt ST, Engel SR; SGD Project. An Introduction to the Saccharomyces Genome Database (SGD). Methods Mol Biol. 2018;1757:21–30. doi:10.1007/978-1-4939-7737-6_2.

Lee R, Sternberg P. Building a cell and anatomy ontology of Caenorhabditis elegans. Comp Funct Genomics. 2003; 4:121–126. doi:10.1002/cfg.248

Lydeard JR, Lipkin-Moore Z, Sheu Y-J, Stillman B, Burgers PM, Haber JE. Break-induced replication requires all essential DNA replication factors except those specific for pre-RC assembly. Genes Dev. 2010;24(11):1133–1144. doi:10.1101/gad.1922610.

MacDougall A, Volynkin V, Saidi R, Poggioli D, Zellner H, Hatton-Ellis E, Joshi V, O'Donovan C, Orchard S, Auchincloss AH, *et al.* Unirule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics. 2021;36(22–23):5562. doi:10.1093/bioinformatics/btaa663.

McDonald AG, Tipton KF. Fifty-five years of enzyme classification: advances and difficulties. FEBS J. 2014;281(2):583–592. doi:10.1111/febs.12530.

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

McIntosh BK, Renfro DP, Knapp GS, Lairikyengbam CR, Liles NM, Niu L, Supak AM, Venkatraman A, Zweifel AE, Siegele DA, *et al.* Ecoliwiki: a wiki-based community resource for Escherichia coli. Nucleic Acids Res. 2012;40(D1):D1270–7. doi:10.1093/nar/gkr880.

Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horáčková A, Melicher F, Perfetto L, Pokorný D, Lopez MR, Türková A, *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res 2019; 47(D1):D550–D558. doi:10.1093/nar/gky1001.

Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14(3):703–721. doi:10.1038/s41596-019-0128-8.

Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. J Biomed Inform. 2011;44(1): 87–93. doi:10.1016/j.jbi.2010.03.002.

Mungall CJ, Torniai C, Gkoutos GV, Lewis SL, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012; 13:R5. doi:10.1186/gb-2012-13-1-r5

Munoz-Torres M, Carbon S. Get GO! Retrieving GO data using AmiGO, QuickGO, API, files, and tools. Methods Mol Biol. 2017; 1446:149–160. doi:10.1007/978-1-4939-3743-1_11.

Nadendla S, Jackson R, Munro J, Quaglia F, Mészáros B, Olley D, Hobbs ET, Goralski SM, Chibucos M, Mungall CJ, *et al.* ECO: the Evidence and Conclusion Ontology, an update for 2022. Nucleic Acids Res. 2022;50(D1):D1515–D1521. doi:10.1093/nar/gkab1025.

Natale DA, Arighi CN, Blake JA, Bona J, Chen C, Chen SC, Christie KR, Cowart J, D'Eustachio P, Diehl AD, *et al.* Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. Nucleic Acids Res. 2017;45(D1):D339–D346. doi:10.1093/nar/gkw1075.

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, *et al.* Interpro in 2022. Nucleic Acids Res. 2022;51(D1):D418–D427. doi:10.1093/nar/gkac993.

Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, Pajkos M, Lazar T, Peña-Díaz S, Santos J, *et al.* Disprot in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Res. 2022;50(D1):D480–D487. doi:10.1093/nar/gkab1082.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, *et al.* A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–227. doi:10.1038/nmeth.2340.

Ramsey J, McIntosh B, Renfro D, Aleksander SA, LaBonte S, Ross C, Zweifel AE, Liles N, Farrar S, Gill JJ, *et al.* Crowdsourcing biocuration: The Community Assessment of Community Annotation with Ontologies (CACAO). PLoS Comput Biol. 2021;17(10): e1009463. doi:10.1371/journal.pcbi.1009463.

Saverimuttu SCC, Kramarz B, Rodríguez-López M, Garmiri P, Attrill H, Thurlow KE, Makris M, de Miranda Pinheiro S, Orchard S, Lovering RC. Gene Ontology curation of the blood-brain barrier to improve the analysis of Alzheimer's And other neurological diseases. Database. 2021;2021:baab067. doi:10.1093/database/baab067.

Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, *et al.* Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2020;48(D1):D9–16. doi:10.1093/nar/gkz899.

Sian L, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, *et al.* Flybase: a guided tour of highlighted features. Genetics. 2022; 220:iyac035. doi:10.1093/genetics/iyac035.

Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45(D1): D592–D596. doi:10.1093/nar/gkw924.

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46. doi:10.1186/gb-2005-6-5-r46.

Smith JR, Hayman GT, Wang S-J, Laulederkind SJF, Hoffman MJ, Kaldunski ML, Tutaj M, Thota J, Nalabolu HS, Ellanki SLR, *et al.* The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. Nucleic Acids Res. 2020;48:D731–D742. doi:10.1093/nar/gkz1041.

Thomas PD. The Gene Ontology and the meaning of biological function. Methods Mol Biol. 2017;1446:15–24. doi:10.1007/978-1-4939-3743-1_2.

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: making genome-scale phylogenetics accessible to all. Protein Sci. 2022;31(1):8–22. doi:10.1002/pro.4218.

Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, Balhoff JP, Albou L-P, Good B, Gaudet P, *et al.* Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019;51(10):1429–1433. doi:10.1038/s41588-019-0500-1.

Thompson W, Zuk RT. Acylation of CDP-monoacylglycerol cannot be confirmed. J Biol Chem. 1983;258(16):9623. doi:10.1016/S0021-9258(17)44541-8.

Torto-Alalibo T, Collmer CW, Gwinn-Giglio M. The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. BMC Microbiol. 2009;9 Suppl1:S1. doi:10.1186/1471-2180-9-S1-S1

The UniProt Consortium. Uniprot: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–D169. doi:10.1093/nar/gkw1099.

Walls RL, Cooper L, Elser J, Gandolfo MA, Mungall CJ, Smith B, Stevenson DW, Jaiswal P. The Plant Ontology facilitates comparisons of plant development stages across species. Front Plant Sci. 2019;10:631. doi:10.3389/fpls.2019.00631.

Winsor GL, Lo R, Ho Sui SJ, Ung KS, Huang S, Cheng D, Ching WK, Hancock RE, Brinkman FS. Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. Nucleic Acids Res. 2005; 33(Database issue):D338–D343. doi:10.1093/nar/gki047.

Editor: A. Baryshnikova

Downloaded from academic.oup.com/genetics/article/224/1/iyad031/7068118 by Kungl Tekniska Hogskolan Biblioteket user on 02 October 2026

A.3 Evolution

Recommended reading for Chapter 3 (listed in Section 3.9).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Kimura (1968) — the neutral theory in two pages; why most molecular change may be neutral.
  DOI · via SU library (SU login)

  Evolutionary Rate at the Molecular Level. Kimura (1968). *Nature*. doi:10.1038/217624a0

- Koonin (2005) — orthologues, paralogues and evolutionary genomics; the definitions and why they matter.
  DOI · via SU library (SU login)

  Orthologs, Paralogs, and Evolutionary Genomics. Koonin (2005). *Annual Review of Genetics*. doi:10.1146/annurev.genet.39.073003.114725

  > Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Orthology and paralogy are key concepts of evolutionary genomics. A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Genome comparisons show that orthologous relationships with genes from taxonomically distant species can be established for the majority of the genes from each sequenced genome. This review examines in depth the definitions and subtypes of orthologs and paralogs, outlines the principal methodological approaches employed for identification of orthology and paralogy, and considers evolutionary and functional implications of these concepts.

- Lynch and Conery (2000) — how often genes duplicate and what happens to the copies.
  DOI · via SU library (SU login)

  The Evolutionary Fate and Consequences of Duplicate Genes. Lynch & Conery (2000). *Science*. doi:10.1126/science.290.5494.1151

  > Gene duplication has generally been viewed as a necessary source of material for the origin of evolutionary novelties, but it is unclear how often gene duplicates arise and how frequently they evolve new functions. Observations from the genomic databases for several eukaryotic species suggest that duplicate genes arise at a very high rate, on average 0.01 per gene per million years. Most duplicated genes experience a brief period of relaxed selection early in their history, with a moderate fraction of them evolving in an effectively neutral manner during this period. However, the vast majority of gene duplicates are silenced within a few million years, with the few survivors subsequently experiencing strong purifying selection. Although duplicate genes may only rarely evolve new functions, the stochastic silencing of such genes may play a significant role in the passive origin of new species.

- Echave et al. (2016) — why rates of evolution vary among protein sites; the link between structure and evolution.
  DOI · via SU library (SU login)

  Causes of evolutionary rate variation among protein sites. Echave et al. (2016). *Nature Reviews Genetics*. doi:10.1038/nrg.2015.18

  > It has long been recognized that certain sites within a protein, such as sites in the protein core or catalytic residues in enzymes, are evolutionarily more conserved than other sites. However, our understanding of rate variation among sites remains surprisingly limited. Recent progress to address this includes the development of a wide array of reliable methods to estimate site-specific substitution rates from sequence alignments. In addition, several molecular traits have been identified that correlate with site-specific mutation rates, and novel mechanistic biophysical models have been proposed to explain the observed correlations. Nonetheless, current models explain, at best, approximately 60% of the observed variance, highlighting the limitations of current methods and models and the need for new research directions.

- Treangen and Rocha (2011) — horizontal transfer, not duplication, drives gene family expansion in prokaryotes.
  full text below · DOI · open access (CC BY 4.0)

  Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. Treangen & Rocha (2011). *PLoS Genetics*. doi:10.1371/journal.pgen.1001284

  > Gene duplication followed by neo- or sub-functionalization deeply impacts the evolution of protein families and is regarded as the main source of adaptive functional novelty in eukaryotes. While there is ample evidence of adaptive gene duplication in prokaryotes, it is not clear whether duplication outweighs the contribution of horizontal gene transfer in the expansion of protein families. We analyzed closely related prokaryote strains or species with small genomes (Helicobacter, Neisseria, Streptococcus, Sulfolobus), average-sized genomes (Bacillus, Enterobacteriaceae), and large genomes (Pseudomonas, Bradyrhizobiaceae) to untangle the effects of duplication and horizontal transfer. After removing the effects of transposable elements and phages, we show that the vast majority of expansions of protein families are due to transfer, even among large genomes. Transferred genes--xenologs--persist longer in prokaryotic lineages possibly due to a higher/longer adaptive role. On the other hand, duplicated genes--paralogs--are expressed more, and, when persistent, they evolve slower. This suggests that gene transfer and gene duplication have very different roles in shaping the evolution of biological systems: transfer allows the acquisition of new functions and duplication leads to higher gene dosage. Accordingly, we show that paralogs share most protein-protein interactions and genetic regulators, whereas xenologs share very few of them. Prokaryotes invented most of life's biochemical diversity. Therefore, the study of the evolution of biology systems should explicitly account for the predominant role of horizontal gene transfer in the diversification of protein families.

- Emms and Kelly (2019) — OrthoFinder; how orthology is inferred from gene trees at genome scale.
  full text below · DOI · open access (CC BY 4.0)

  OrthoFinder: phylogenetic orthology inference for comparative genomics. Emms & Kelly (2019). *Genome Biology*. doi:10.1186/s13059-019-1832-y

  > Here, we present a major advance of the OrthoFinder method. This extends OrthoFinder's high accuracy orthogroup inference to provide phylogenetic inference of orthologs, rooted gene trees, gene duplication events, the rooted species tree, and comparative genomics statistics. Each output is benchmarked on appropriate real or simulated datasets, and where comparable methods exist, OrthoFinder is equivalent to or outperforms these methods. Furthermore, OrthoFinder is the most accurate ortholog inference method on the Quest for Orthologs benchmark test. Finally, OrthoFinder's comprehensive phylogenetic analysis is achieved with equivalent speed and scalability to the fastest, score-based heuristic methods. OrthoFinder is available at https://github.com/davidemms/OrthoFinder.

The open-access papers are reproduced below under their original licences.

OPEN ACCESS Freely available online

PLoS GENETICS

Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes

Todd J. Treangen[1,2,3¤]*, Eduardo P. C. Rocha[1,2,3]

1 Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, Paris, France, **2** CNRS, URA2171, Paris, France, **3** UPMC Université Pierre et Marie Curie, Atelier de Bioinformatique, Paris, France

Abstract

Gene duplication followed by neo- or sub-functionalization deeply impacts the evolution of protein families and is regarded as the main source of adaptive functional novelty in eukaryotes. While there is ample evidence of adaptive gene duplication in prokaryotes, it is not clear whether duplication outweighs the contribution of horizontal gene transfer in the expansion of protein families. We analyzed closely related prokaryote strains or species with small genomes (*Helicobacter, Neisseria, Streptococcus, Sulfolobus*), average-sized genomes (*Bacillus*, Enterobacteriaceae), and large genomes (*Pseudomonas*, Bradyrhizobiaceae) to untangle the effects of duplication and horizontal transfer. After removing the effects of transposable elements and phages, we show that the vast majority of expansions of protein families are due to transfer, even among large genomes. Transferred genes—xenologs—persist longer in prokaryotic lineages possibly due to a higher/longer adaptive role. On the other hand, duplicated genes—paralogs—are expressed more, and, when persistent, they evolve slower. This suggests that gene transfer and gene duplication have very different roles in shaping the evolution of biological systems: transfer allows the acquisition of new functions and duplication leads to higher gene dosage. Accordingly, we show that paralogs share most protein–protein interactions and genetic regulators, whereas xenologs share very few of them. Prokaryotes invented most of life's biochemical diversity. Therefore, the study of the evolution of biology systems should explicitly account for the predominant role of horizontal gene transfer in the diversification of protein families.

Citation: Treangen TJ, Rocha EPC (2011) Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLoS Genet 7(1): e1001284. doi:10.1371/journal.pgen.1001284

Editor: Nancy A. Moran, Yale University, United States of America

Received June 14, 2010; **Accepted** December 20, 2010; **Published** January 27, 2011

Copyright: © 2011 Treangen, Rocha. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was carried out with financial assistance from the Institut Pasteur and the Centre National de la Recherche Scientifique. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: treangen@pasteur.fr

¤ Current address: Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America

Introduction

Prokaryotes have highly variable gene repertoires, varying from just over 100 genes to nearly 10000 [1,2]. Such variations in genome size are typically associated with expansions and contractions of protein families. Expansions of protein families are associated with the acquisition of novel functions, novel regulatory structures and system robustness [3–5]. They can take place by horizontal gene transfer, in which case homologs are called xenologs, or by intra-chromosomal duplication, in which case homologs are called paralogs. Genes are identical upon duplication and if there is no direct selection on multiple copies, the redundant extra copies are quickly lost. Yet, sometimes gene duplications persist in genomes due to selection for increased gene dosage. Periods of such adaptive gene duplication are windows of opportunity for the acquisition of new functions by the slow evolutionary divergence between the duplicates [6–8]. Duplicated genes can thus be fixed because sub- or neo-functionalization processes render the two copies adaptive when selection for higher gene dosage ceases (reviewed in [9–12]). There is ample evidence that intra-chromosomal gene duplication (IGD) has an adaptive role in bacteria [13,14], e.g. in antigenic variation, antibiotic resistance or in genome expansion [8,15–19]. Furthermore, models aiming at explaining the patterns of biological networks are in general based on the preconception that protein families expand by gene duplication processes [20,21].

Horizontal gene transfer (HGT) results in the acquisition of radically new genetic information, liberating genomic evolutionary processes from tinkering exclusively with pre-existing genes [22–24]. HGT occurs at very high rates leading to very large species pan-genomes [25]. Yet, while up to 96% of the genes in a given prokaryote genome might have been affected by HGT [26–28], it has been estimated that HGT contributes at best to 25% of all expansions of protein families [26,29–32] (22% overall but 60% for large protein families in [23]). These contradictory conclusions require an explanation. Prior analyses on the relative abundance of paralogs over xenologs were performed several years ago using the available distant genomes. However, comparisons of genomes from very divergent lineages pose several problems: (i) disambiguation of orthologs, paralogs and xenologs is very error-prone, (ii) lineages are saturated with changes, (iii) expansions within terminal branches are associated with paralogy when they can result from HGT between closely related taxa. In fact, long terminal branches hide a complex history of duplications and transfer and consequently the relative roles of transfer and duplication can only be assessed using *closely* related genomes.

Here, we apply a protein family analysis pipeline to a multitude of closely related complete genomes to re-examine the role of

Author Summary

Prokaryotes can be found in the most diverse and severe ecological niches of the planet. Their rapid adaptation is, in part, the result of the ability to acquire genetic information horizontally. This means that prokaryotes utilize two major paths to expand their repertoire of protein families: they can duplicate a pre-existing gene or acquire it by horizontal transfer. In this study, we track family expansions among closely related strains of prokaryotic species. We find that the majority of gene expansions arrive via transfer not via duplication. Additionally, we find that duplicate genes tend be more transient and evolve slower than transferred ones, highlighting different roles with respect to adaptation and evolution. These results suggest that prevailing theories aimed at understanding the evolution of biological systems grounded on gene duplication might be poorly fit to explain the evolution of prokaryotic systems, which include the vast majority of life's biochemical diversity.

transfer and duplication in expansion of protein families in prokaryotes. Our comparisons involve genomes that have diverged recently (Figures S1 and S2). As a result, we can use sequence similarity and co-localization information to infer accurate core genomes, phylogenies and ancestral events. At these evolutionary distances, paralogs arising from duplications having occurred since the last common ancestor are very similar in sequence and thus can easily be separated from most xenologs. Finally, at these evolutionary scales the population data enables the analysis of fixation of protein family expansions and test if some processes have a longer-term effect on genomes than others.

Results

Identification and characterization of protein families

We analyzed 110 genomes encompassing eight distant clades (Table S1): *Helicobacter, Neisseria, Streptococcus and Sulfolobus* (small genomes, ~2 Mb), Enterobacteriaceae and *Bacillus* (average, ~4–5 Mb), Bradyrhizobiaceae *and Pseudomonas* (large, ~6 Mb). The clades were selected to include a large number of closely related genomes and to span diverse phyla, G+C contents and genome sizes (Table 1). By using sequence similarity analysis followed by clustering (see Materials and Methods), we obtained a set of 59,541 families containing a total of 419,035 proteins (Table 1). Families with exactly one member per genome and where all members are highly similar (*i.e.* putative orthologs) were used to define the core genomes and to build highly robust phylogenetic trees of the clades (100% bootstraps in nearly all branches, Figure S3). Many families (48%) consisted of singletons, i.e. they contained one single protein, thereby confirming that new families in prokaryotes are introduced at high rates by HGT. These families were excluded from the analysis.

Given the pattern of presence/absence of genes and a reference tree, we inferred gene expansion events with BayesTraits [33]. Families without protein expansions within the lineages were excluded from further analysis. We also removed IS and phage sequences, but not the corresponding cargo regions, because these elements constitute a large fraction of repeats in genomes [15], are known to be horizontally transferred and are in general quickly lost [34,35]. Our main analysis includes the remaining 3190 families. These families have few members (0–3) in each genome (Figure 1) showing that expansions are rare at these narrow evolutionary scales. This is in agreement with high rates of change but low rates of fixation of changes in gene repertoires [36,37]. The most frequent event found in our data was the gain of a paralog or xenolog by the largest genomes in the clade (Table S2), such as *S. agalactiae* NEM316 (20% of all expansions in *Streptococcus*). Around 50% of all gene expansions occur in the clade with largest genomes –*Bradyrhizobium* - with 35% occurring in the two largest genomes of that clade. Although families with recent expansions account for a small fraction of genomes this is in good agreement with an association between expansions of protein families and increased genome size.

Expansions of protein families arise most frequently by HGT

Duplication processes in prokaryotes produce tandem, strictly identical copies of genes [8,38,39]. At the evolutionary distances considered in this work, this implies that paralogs arising in the lineages are co-localized because of the low rearrangement rates and are highly similar in sequence because of the low mutation rates. On the other hand, transferred genes are inserted almost randomly in genomes and show a large range of sequence similarity relative to the native homolog. Therefore, we disentangle IGD from HGT using sequence similarity and positional information (see Materials and Methods). We define a minimal similarity threshold for paralogs assuming that they evolve at rates close to the ones of the core genome. This threshold is therefore based on the distribution of similarities between pairs of orthologs of the core genome. More precisely, we define three thresholds corresponding to values where protein divergence exceeds that of 95%, 99% or 99.9% of the comparisons between orthologous core genes. Different thresholds produced qualitatively similar results

Table 1. Expansions of gene families.

Clade	genomes	genes per genome	families	families w/expansions	paralogs & xenologs
Enterobacteriaceae	41	4881	15729	1148	6803
Streptococcus	17	1965	4474	317	1255
Bradyrhizobiaceae	9	6363	14231	1009	2203
Helicobacter	8	1525	2446	88	105
Neisseria	11	2120	4922	247	335
Sulfolobus	7	2780	3782	332	505
Bacillus	12	5264	7745	461	1070
Pseudomonas	5	5466	6212	381	405

doi:10.1371/journal.pgen.1001284.t001

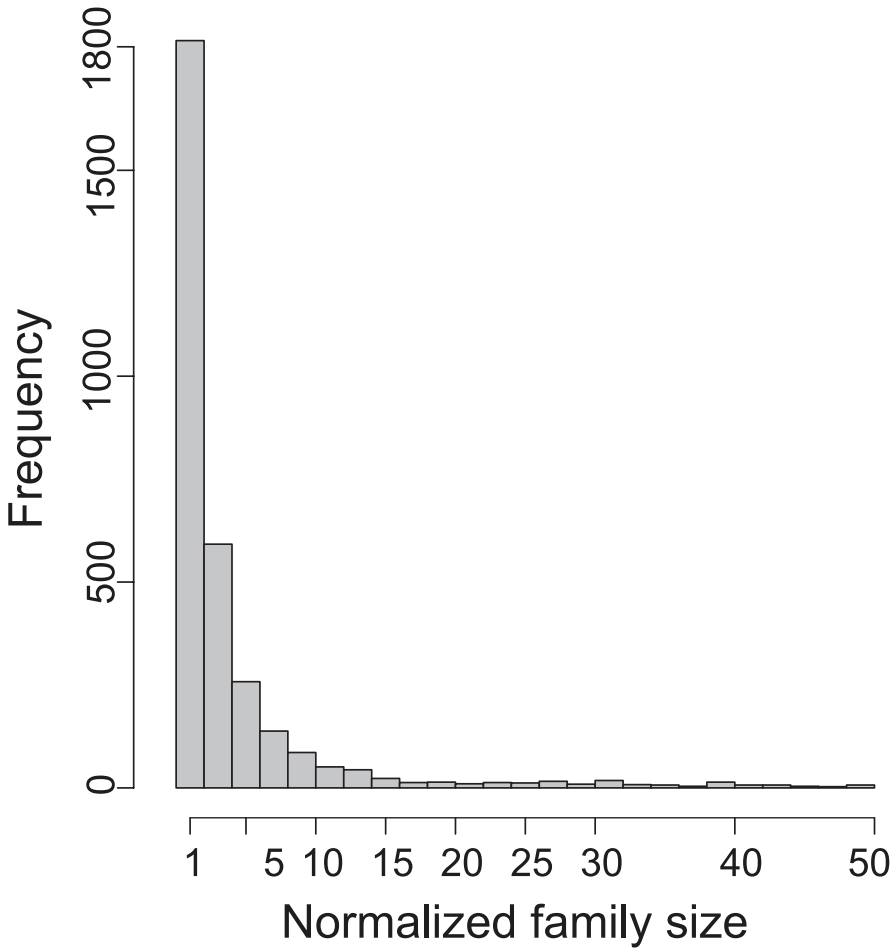


Figure 1. Histogram of the normalized size of gene families. For each family we compute the number of genes in the family and subtract it by the number of genomes containing at least one member of the family.
doi:10.1371/journal.pgen.1001284.g001

and we use the 99% thresholds throughout this work (see Materials and Methods, Figure S4). The use of sequence similarity alone will lead to spurious classification of genes transferred between closely related genomes, because such genes are expected to be highly similar. We thus add the positional criterion considering that two genes are co-localized if no core gene separates them. Expansions producing highly similar co-localized homologs are thus assumed to be created by IGD, whereas the others are created by HGT. We find that the vast majority, between 88% and 98%, of the expansions of protein families are due to HGT (Figure 2). This is true even for large genomes, such as in Bradyrhizobiaceae, which have been proposed to increase in size by gene duplication. Hence, by default, expansions are much more likely to arise by transfer than by internal duplication.

The majority (65%) of the expansions are assigned to HGT using either the criterion of sequence similarity or of co-localization, showing the robustness of the method. If one removes the co-localization criterion, thus assuming that expansions of highly similar proteins arise exclusively by IGD, we still find HGT as the major cause for expansions of protein families in 6 of the 8 clades. The two exceptions, *Neisseria* and *Helicobacter*, have a disproportionate fraction of highly similar homologs in different locations in the genome so that removing the co-localization criterion impacts significantly in the relative contribution of IGD (resp. accounting for 78% and 45%). It is well-established that such repeated elements often engage in gene conversion for antigenic variation both in *Neisseria* and in *Helicobacter* [40,41]. Interestingly, both clades naturally transform conspecific DNA and are known to have extremely high rates of intra-species horizontal transfer [42,43]. The high frequency of HGT within the species is liable to produce a large fraction of very similar xenologs scattered in the genome. This factor is presumably corrected by the use of the co-localization criterion. Importantly, the relative abundance of xenologs over paralogs in the largest genomes, the ones with more expansions, remains almost unchanged when removing the co-localization criterion.

Xenologs persist longer than paralogs

HGT might be less relevant if xenologs were lost at higher rates than paralogs. We calculated the age of introduction of each gene in its lineage. We found that xenologs have an average age of introduction that is twice that of paralogs (Mann-Whitney-Wilcoxon, $p<0.0001$). Xenologs are therefore more persistent than paralogs (Table 2). These results remain qualitatively unchanged when discarding positional information (same test, $p<0.0001$). This trend holds for all clades, except for *Helicobacter* and Bradyrhizobiaceae, where differences are not statistically significant. The phylogenetic tree of *Helicobacter* has very small

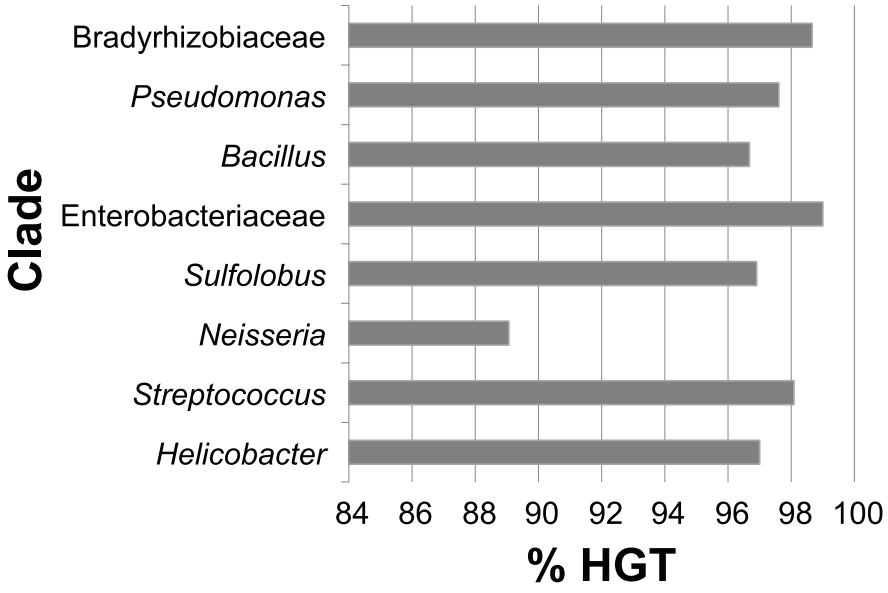


Figure 2. Relative contribution of horizontal gene transfer in protein family expansions.
doi:10.1371/journal.pgen.1001284.g002

internal branches (Figure S3) and we only identified 25 ancient events in them. All 25 ancient events correspond to HGT suggesting that paralogs are quickly purged also in *Helicobacter*. In Bradyrhizobiaceae all terminal branches are long thereby reflecting a mixture of recent and ancient events (Figure S2). There is thus no significant difference between internal and terminal nodes. Yet, HGT created ~98% of family expansions in both the internal and the terminal branches of this lineage. Overall, these results show that expansions by HGT are more abundant and remain for longer periods of time in genomes.

Mobile elements spuriously inflate the estimates of paralogy

Prophages and IS were removed from the previous analyses. Since they represent 10–40% of all expansions of protein families (Figure 3) we analyzed the effect of re-introducing them. As expected this resulted in similar conclusions. When including IS and phages in the analyses we still find that HGT accounts for more than 88% of expansions in all clades. However, if we include prophages and IS and drop the co-localization criterion, we infer a dramatic increase in duplications (Figure 3). The effect is most pronounced in Enterobacteriaceae where expansions associated to IGD increase 100 fold, largely caused by the abundance of homologous lambda-like prophages in these genomes and at a much lesser extent by the expansion of transposable elements in *Shigella*. In spite of this, expansions of protein families are predominantly classified as HGT in all but the smallest genomes (*Neisseria*, *Streptococcus*, *Sulfolobus*). The scattered insertion of homologous prophage and transposable elements in genomes can thus introduce in the analysis many family expansions that are spuriously assigned to gene duplications when in fact they result from horizontal transfer and are quickly lost. The co-localization criterion renders the method robust to mobile elements even when they are not explicitly removed. In the subsequent analyses we use the original dataset where prophages and IS are excluded.

Table 2. Different (Δ) ages of IGD and HGT per clade.

	With positional information			Without[1]
	HGT	IGD	Δ [2]	Δ [2]
Streptococcus	0.10	0.0	+0.10*	+0.09**
Enterobacteriaceae	0.03	0.0	+0.03**	+0.03**
Bradyrhizobiaceae	0.10	0.10	0.00	+0.01**
Bacillus	0.05	0.02	+0.03**	+0.01**
Pseudomonas	0.04	0.01	+0.03*	+0.03**
Helicobacter	0.04	0.04	0.00	+0.02**
Neisseria	0.03	0.02	+0.01*	+0.01**
Sulfolobus	0.04	0.0	+0.04**	+0.02**

1 Analysis using only the sequence threshold to disambiguate between paralogs and xenologs.
2 Mann-Whitney-Wilcoxon test:
*$p<0.01$,
**<0.001.
doi:10.1371/journal.pgen.1001284.t002

Paralogs are more highly expressed

Are there significant differences between protein families expanding by duplication or transfer? To answer this question we restricted the following analyses to *E. coli* because it is the only species for which there is sufficient genome-wide experimental data. We first tested if the two classes of protein families have different gene expression levels using the codon adaptation index (CAI; [44]). CAI is higher for paralogs than for xenologs (0.68 and 0.48, Mann-Whitney-Wilcoxon test, $p<0.0001$) and smallest for genes in families without expansions (0.42, Figure 4). Similar results were found using proteomic data (emPAI) [45] ($p<0.005$). We conclude that paralogs are more expressed than xenologs. While this is at odds with previous work on *E. coli*, where genes with the lowest CAI were found to be overrepresented among recent duplications [30], this discrepancy probably results from our explicit removal of IS and phages that have very low CAI because they are A+T rich and are lowly expressed under exponential growth. When such elements are removed from the

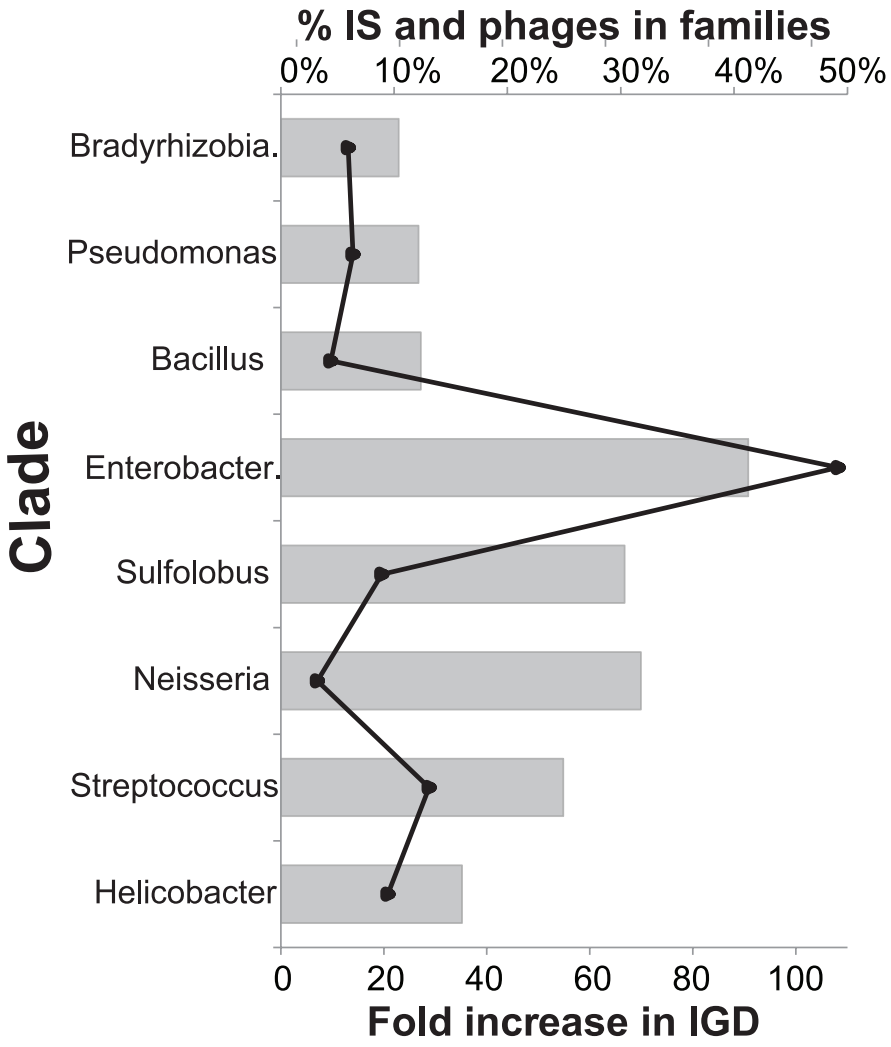


Figure 3. Abundance of IS and prophages and increased inference of IGD events when included in analysis. The bar plot (left y-axis) shows the percentage of gene family expansions of IS and phage origin. The line plot (right y-axis) indicates the increase of the number of expansions assigned to duplications when the co-localization criterion is ignored and IS and prophages are included in the dataset.
doi:10.1371/journal.pgen.1001284.g003

analysis it is clear that families with expansions are more expressed than average and, in particular, the families producing paralogs.

Xenologs evolve faster

To analyze the evolutionary rates of paralogs and xenologs we aligned them to the corresponding ortholog, when there was one. We analyzed separately non-synonymous (dN) and synonymous (dS) substitution rates, after excluding highly divergent pairs (if $dS>1.5$) [46]. As expected we found the highest dN and dS values in xenologs ($dS_{HGT}=0.43$, $dS_{IGD}=0.26$, $p<0.001$; $dN_{HGT}=0.07$, $dN_{IGD}=0.03$, $p<0.001$, Mann-Whitney-Wilcoxon tests). The lower values of dN and dS reflect the recent divergence of paralogs relative to xenologs and the methodological bias of imposing a sequence similarity filter to paralogs. But the plot of dS against dN shows that dN/dS is 50% higher for xenologs than for paralogs over the entire range of dN and dS values ($p=0.01$, Wilcoxon test, Figure 5). This strongly suggests that xenologs evolve faster. Yet, we capped paralog similarity in proteins using the sequence similarity threshold and this might bias our dN/dS estimate. To make a direct comparison of dS and dN values between paralogs and xenologs, we filtered the latter using the same similarity threshold as used for paralogs. We found that both dN and dS were higher in xenologs than in paralogs ($p<0.01$, Mann-Whitney-Wilcoxon tests). We also tested if higher evolutionary rates in xenologs were caused by amelioration to the host GC content [47] (see Materials and Methods). As expected, xenologs have higher G+C deviations to the core genome than paralogs (0.05 vs. 0.04, $p<0.01$, Wilcoxon test). This difference is small, as found previously in *Salmonella* [48], possibly because we removed IS and phages. For similar GC deviations we still found that xenologs have higher synonymous and non-synonymous substitution rates (Mann-Whitney-Wilcoxon tests on the ratio of the rate over Δ(G+C), both $p<0.05$). These results show that xenologs evolve faster. This could be caused by relaxed selection on protein function in xenologs relative to paralogs, but this explanation seems at odds with our observation that xenologs are more persistent than paralogs. Instead, the results are consistent with the observed negative correlation between expression levels and both dN and dS in *E. coli* [49]. Paralogs being more expressed than xenologs, they evolve slower in synonymous positions, due to

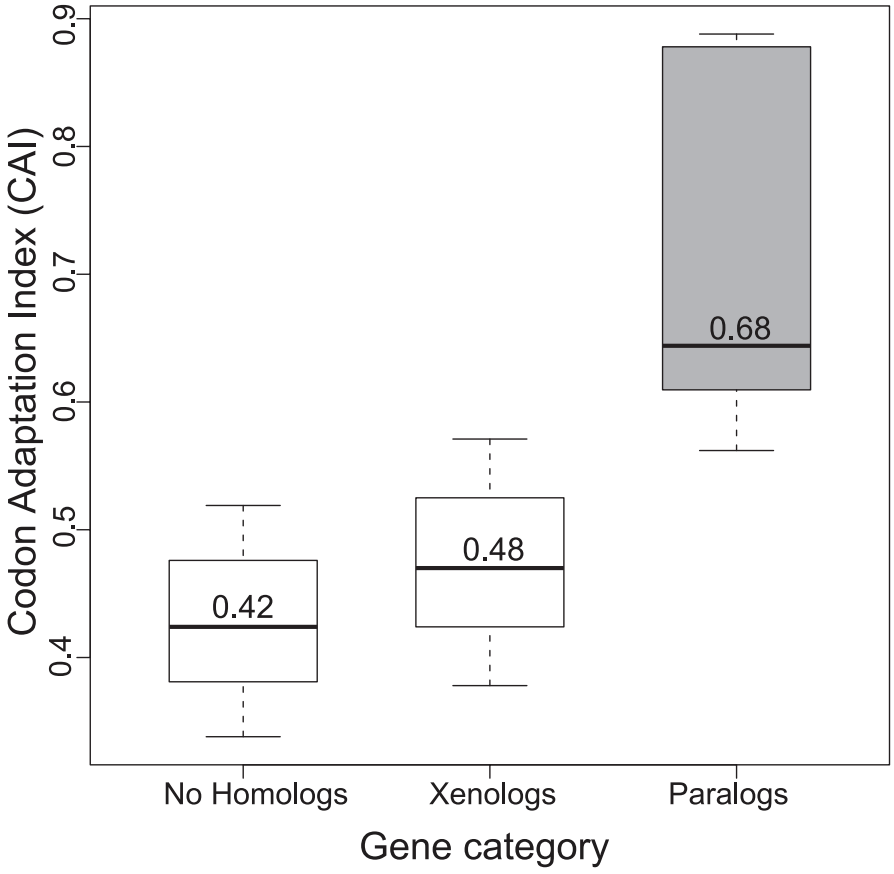


Figure 4. Gene expression differs according to gene origin. Paralogs are more expressed, as measured by the codon adaptation index, than xenologs. Xenologs, however, are more expressed than the genes without paralogs and xenologs.
doi:10.1371/journal.pgen.1001284.g004

selection on codon usage bias, and even slower in non-synonymous positions, due to purifying selection on protein sequence.

Xenologs share fewer protein interactions

To assess the differences between the processes of protein family expansions in terms of cellular networks we used a set of 74,776 protein-protein interactions (PPI) in *E. coli* MG1655 [50]. We computed the fraction of interactions shared by pairs of proteins in families with expansions (ΔPPI). As expected, paralogs share more interactions than xenologs (ΔPPI 1.0 and 0.30, $p<0.01$, Mann-Whitney-Wilcoxon test), even when discarding positional information in paralog and xenolog disambiguation (0.61 and 0.29, $p<0.001$, same test). As similarity increases between the pair of proteins so does the percentage of shared protein interactions ($R^2 = 0.60$, $p<0.01$). Highly similar xenologs have ΔPPI similar to paralogs. Hence, the differences observed between xenologs and paralogs in terms of their integration in protein-protein interactions reflect the intrinsic consequences of the mechanisms of acquisition; xenologs may or may not be very similar in sequence, but recent paralogs are necessarily very similar. As a result, PPI networks will evolve very differently in lineages affected by horizontal transfer relative to lineages with strictly vertical transmission of genetic information.

Xenologs share fewer regulatory mechanisms

We then estimated the genetic regulation overlap between paralogs and between xenologs. We computed the fraction of shared regulatory interactions as the number of regulators interacting with both homologs over the sum of regulators interacting with at least one of them. Homologs with no known interactions were recorded as zero. We used *E. coli* data [51] on 2020 regulators including regulation by (1) transcription factors, (2) RNA-binding proteins, (3) sigma factors, (4) protein–protein interactions and 5) DNA supercoiling. As expected, paralogs share more regulators than xenologs (0.1 to 0.0, $p = 0.06$, Mann-Whitney-Wilcoxon test). The difference is small because most genes have no known regulators in contemporary datasets. We therefore restricted our analysis to genes known to be under the control of at least one regulator. In this case, paralogs share twice the regulators of xenologs ($p<0.0001$, same test). Hence, the evolution of regulatory networks by expansion of protein families depends on the type of mechanism involved. As it is the case for protein-protein interactions, duplications generate paralogs with overlapping regulatory dependencies, whereas transfer generates xenologs that potentially have highly different regulatory dependencies. This picture is in agreement with the frequently observed co-transfer of transcription factors and their target genes [52], and subsequent slow integration of these sub-networks in the larger genetic network [53].

Discussion

Contrary to previous studies, we found a high rate of HGT in all eight clades of prokaryotes analyzed in this study. At least 88% of all

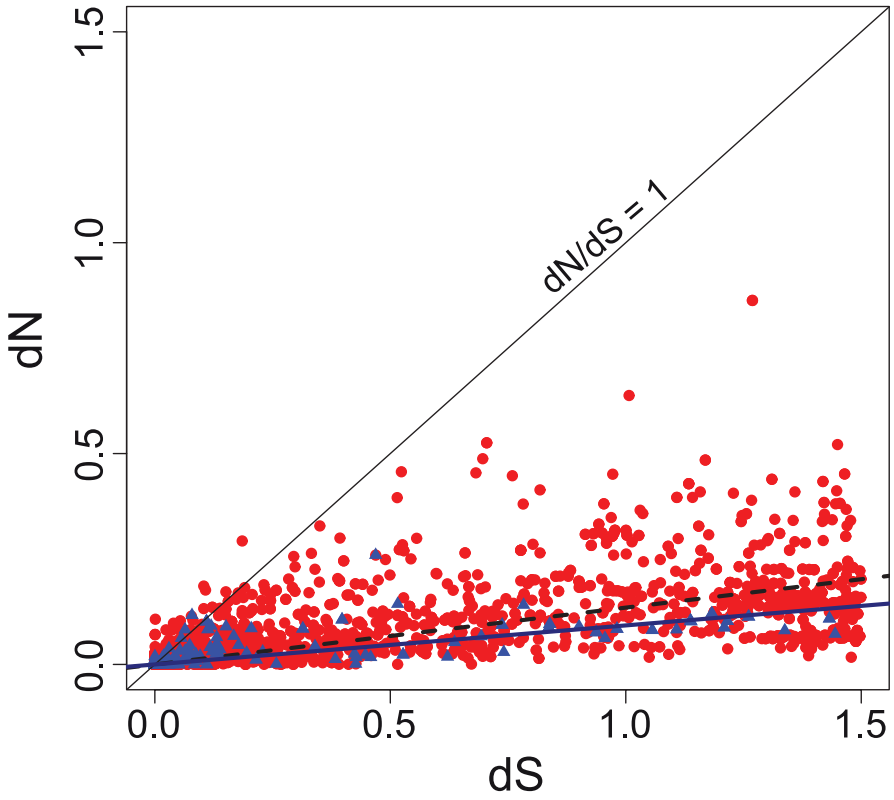


Figure 5. Evolutionary rates differ between paralogs and xenologs. Non-synonymous (dN) and synonymous (dS) substitution rates in paralogs (blue; dashed linear fit) and xenologs (red; solid linear fit) in all clades computed using Codeml from PAML [76] (model = 1, fix_omega = 0).
doi:10.1371/journal.pgen.1001284.g005

expansions of protein families originated by HGT noth when including or discarding prophages and IS. The largest genomes in each clade are also the ones containing more new xenologs. This is concordant with the proposed role of horizontal transfer in the increase in size of the largest bacterial genomes [54]. The important contribution of xenology to family expansions is also consistent with data showing *E. coli*'s promiscuity to most transferred genes [55] and data highlighting the impact of transferred genes on the evolution of *Salmonella* [56]. It has been proposed that due to deletion biases in prokaryotes, gene duplications are not afforded with sufficient opportunity for neo- or sub-functionalization, reducing its role to transient gene amplification [57]. Xenologs are related, yet different, proteins that can provide a potentially advantageous distinct function immediately upon transfer. Since they tend to be located apart from the native homolog they are also less prone to genetic deletion. These two effects contribute to explain the higher persistence of xenologs.

Our disambiguation method uses information on gene co-localization and protein sequence similarity, which can both lead to misclassifications. There are two major sources of errors when using the co-localization criterion. First, while amplifications generally occur in tandem, they may be separated by genome rearrangements and produce spurious xenologs. This is unlikely to produce many false classifications in the present work because: (i) dropping the co-localization criterion still results in a preponderance of xenologs, (ii) rearrangements are rare at these evolutionary distances [58]. Second, xenologs can integrate close to their homolog in the genome. This can occur by chance alone, as a rare event, or because transferred genes tend to insert in integration hotspots. Additionally, amplified regions in one chromosome can be transferred as one single block into another genome. This will lead to spurious classification of these xenologs as paralogs. We found that around 7% of the highly divergent xenologs are co-localized. It is therefore likely that around 7% of the highly similar xenologs are mis-classified as paralogs. Highly similar xenologs are a minority of all xenologs suggesting that the analysis is not affected in an important way by co-localized gene transfer. On the other hand, the co-localization criterion is very useful to control for two effects that lead to the misclassification of xenologs: (i) the propagation of mobile genetic elements such as prophages, (ii) extensive gene transfer within the species.

The sequence similarity threshold allows classing the highly divergent pairs of genes as xenologs but can be misleading in some circumstances. If paralogs and core genome proteins evolve at similar rates, then the threshold (similarity lower than 99% of genes of the core genome) means that ~1% of duplicated genes are spuriously classed as xenologs. This fraction of false positives has a negligible impact in the quantification of xenology. Yet, our estimation of false positives might be affected by two factors of opposite effect. First, core genes might evolve slower than families with paralogs, placing the sequence similarity threshold too high and thereby increasing the number of spurious xenologs. This is unlikely because paralogs are highly expressed, and thus highly conserved, and high similarity and physical proximity between paralogs favors gene conversion, which also slows evolutionary rates. Second, we find that most expansions are recent, thus duplicates should be very similar, but we use a sequence similarity threshold that includes comparisons involving core genes between all, including the most distant, genomes in the clade. This leads to a conservative (i.e. too low) sequence similarity threshold.

Importantly, variations in the sequence similarity threshold produced no significant changes in the results: *xenolog expansions vastly outnumber paralog expansions*.

Many models assume that duplication is the major mechanism underlying the evolution of protein-protein interactions (PPI) [4,20] or regulatory networks [21]. Duplications constrain genomic evolutionary processes to tinker with pre-existing information producing identical genes that are functionally and genetically redundant. Instead, horizontal transfer allows the acquisition of xenologs with differentiated functions and regulatory networks. These processes have thus different roles in the evolution of protein families. Here, we showed extensive evidence that paralogs and xenologs have different characteristics. Paralogs are more highly expressed, more conserved in sequence and they share regulatory networks and physical interactions. This is consistent with experimental works showing that paralogy is often transiently selected to attain higher gene dosage [8]. Xenologs exhibit expression levels closer to that of the average gene, they differ remarkably from their pair in their physical and genetic interactions and they tend to persist longer in populations. Contrary to the current view, transfer might then be responsible for most long-term adaptive expansions of protein families in prokaryotes. As such, expansions are a particular case of the processes leading to the horizontal transfer of genetic information that is known to shape metabolic [59], genetic [53] and interaction networks [60,61] in prokaryotes. Importantly, recent works have shown that extensive horizontal transfer also exists among eukaryotes [62–64]. These results may thus also be relevant to understand the evolution of biological processes in eukaryotes.

Materials and Methods

Bacterial strains

We analyzed the genome sequences of bacteria from eight widely studied phylogenetic groups representing four different genome sizes (Table S1).

Small (~1.5–3 Mb). 8 *Helicobacter* (7 *Helicobacter* pylori and 1 *Helicobacter acinonychis*), *11 Neisseria* (8 *Neisseria meningitidis, 2 Neisseria gonorrhoeae, 1 Neisseria lactamica*), 17 Streptococcaceae (12 *Streptococcus pyogenes*, 3 *Streptococcus equitans*, *2 Streptococcus* agalactiae). *and 7 Sulfolobus* (6 *Sulfolobus islandicus, 1 Sulfolobus solfataricus*).

Average (4–5 Mb). *41* Enterobacteriaceae (28 *Escherichia and* 13 *Salmonella*), *12 Bacillus* (8 *Bacillus cereus*, 2 *Bacillus thuringiensis*, 1 *Bacillus anthracis*, 1 *Bacillus weihenstephanensis*).

Large (~6 Mb). *9* Bradyrhizobiaceae (6 *Rhodopseudomonas* and *3 Bradyrhizobium*), *5 Pseudomonas (4 Pseudomonas putida and* 1 *Pseudomonas fluorescens)*. Unfortunately, a deeply sequenced clade containing only very large genomes (>7 Mb) is not yet available. On the other end of the spectrum, genomes much smaller than *Helicobacter* have few family expansions. All genomes were retrieved from Genbank, except for *Neisseria lactamica* that was provided by the Sanger Center.

Gene family construction

The procedure used for defining gene families is summarized as follows (Figure 6). First, we performed in each clade all-against-all BLASTP [65] comparisons, inter- and intra- genomically. We set the e-value cutoff to 10^{-7}, required the hits to be at least 100 aa in length and the length of the BLASTP hit to span at least 70% of the length of the smallest protein. Also, similar to the procedure in [23] we evaluated the results with a BLASTP bit score threshold equal to 30% of the maximal bit score (*i.e.* a protein matching itself). Since we are interested in multi-gene families (*i.e.* at least one genome has two or more genes), we clustered pairwise BLASTP hits into multiple relationships using mclblastline from

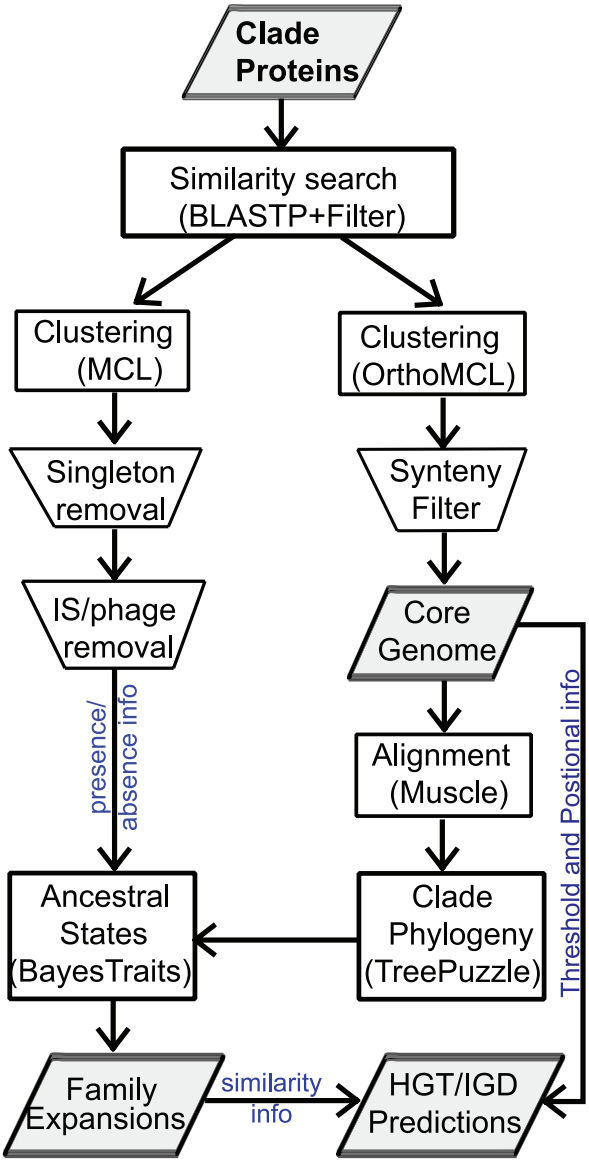


Figure 6. Protein family construction pipeline. Starting with a databank of proteins, we first performed all pairwise similarity searches using BLASTP. The hits were filtered regarding the length of the match (70% of the length of the query) and the bitscore (30% of the maximal bitscore calculated by aligning a protein against itself). To build the gene families we ran MCL blastline and then removed all singletons, IS and Phage. To build the core genome we used OrthoMCL along with a synteny filter based on M-GCAT Clusters. Finally, using presence/absence and phylogenetic information, we obtained the protein families with expansions
doi:10.1371/journal.pgen.1001284.g006

MCL [66], with parameters: **--blast-m9 --blast-score = e --blast-sort = a --mcl-I = 2.0 --mcl-scheme = 7**. After inspection of the results using inflation parameter values from 1.0 to 10.0 in increments of 0.2, we set it to 2.0.

Identification and removal of IS, prophages, and rDNA repeats

Insertion Sequences (IS) were identified by BLASTP query of a custom IS database, including all Uniprot transposases (Touchon,

in preparation) and removed from the analysis. We verified that for *E. coli* our databank contained all IS present in IS Finder [67]. We used Phage_finder to predict prophages [68]. IS are frequently pseudogenized. To remove small fragments of IS elements we searched for regions of homology with known IS in the genome using Repeatoire [69].

Core genome construction

Since genomes in each clade have diverged recently, the orthologs should be highly similar. We enlisted OrthoMCL [70] to build the groups of orthologs inside of the gene families via all-against-all BLASTP comparisons in a clade of interest. OrthoMCL defines putative orthologous relationships between a pair of genomes as BLASTP reciprocal better/best hits. To limit false positives this list was further refined by combining information on the distribution of similarity of these putative orthologs with gene order conservation data (as in [71]). At these distances, gene order conservation is high [58], and positional information can significantly decrease classification errors [72]. Each ortholog pair was then tested for gene order conservation via the ordered list of putative orthologs between the two genomes (i.e. lateral transfer is discounted). Genes not satisfying the constraint are removed from the core genome. Finally, we removed all ortholog pairs less than 65% similar in sequence and differing by more than 30% in length. Naturally, one cannot exclude that some core genes have endured changes by homologous recombination with exogenous genetic material. Yet, gene conversion requires very high similarity and such effects will not affect significantly our study. On the other hand the rare xenologous replacements of core genes by very divergent sequences are expected to lead to loss of synteny and they are therefore discarded from our core genome.

Phylogenetic analyses

The reference phylogenetic tree of each clade was reconstructed from the concatenated alignments of genes comprising their core genome. Protein alignments were generated using MUSCLE [73] and then back-translated to DNA. Tree-Puzzle [74] was used to generate the matrix of distances by maximum likelihood with the HKY+Γ model and exact parameter estimates. The trees were then computed from these distance matrices using BIONJ [75]. We performed 100 bootstrap experiments on the concatenated sequences to assess the robustness of the topology.

Ancestral state reconstruction

We used the reference phylogeny and maximum likelihood (ML) optimization in the "Multistates" component of BayesTraits [33] to estimate ancestral states (0, gene not present, 1, family with no expansions, 2, expanded family, Table S3) for each gene family on all branches of the taxa. We enabled the covarion model for trait evolution, a variant of the continuous-time Markov model that allows for traits to vary their rate of evolution within and between branches. 100 optimization attempts were carried out to find the ML solution. For each node, a gene expansion was considered as present (state = 2) if its probability as estimated by BayesTraits was $\geq$0.5 (the analysis was repeated with p$\geq$0.9 and yielded similar results). In our analysis we only included gene expansion events, not deletions.

Calculating the age of expansions

Using the BayesTraits ancestral state information [33], we calculated the age of expansions as the distance from the leaf node (genome currently containing the paralog or xenolog) to the ancestral node where the gene expansion event occurred (with respect to the root). Expansions appearing in genomes with very long terminal edges were discarded because they might include recent and ancient acquisitions. This includes the following genomes: *E. fergusonii, P. fluorescens, S. equitans, S. agalactiae.*

Classifying the origin of gene family expansions

The key point of our methodology is the disambiguation between paralogs and xenologs by combining two key pieces of evidence: (1) sequence similarity and (2) positional information (Figure 6).

1) Sequence similarity. Genes deriving from a single gene in the last common ancestor of a clade are highly similar because mutations accumulate at a slow pace and the clades selected in this study have diverged recently. In this case the distribution of similarities within gene families of the core genome is narrowly distributed around 100% (Figure 7). As the last common ancestor between two genomes becomes more distant, e.g. comparisons between different species or genera, the range of similarity of genes in the core genome becomes larger. After a certain evolutionary distance, a significant number of duplications preceding the last common ancestor will be less divergent than some orthologs of the core genome because genes evolve at markedly different rates (Figure 7). This is why it is crucial to restrict the analysis to closely related genomes. To define a meaningful threshold of similarity we computed all pairwise sequence similarities between orthologous core genes. We then identified the 99% threshold of sequence similarity, i.e. the sequence similarity below which we found only 1% of the pairwise comparisons. Two genes were regarded as highly similar when their sequence similarity was above this threshold. If the paralogs evolve as the core genome then this threshold implicates that $\sim$1% of paralogs are spuriously assigned as xenologs. We re-did all the analysis with the 95% and 99.9% threshold for the two clades with highest and lowest fraction of xenologs over paralogs (Enterobacteriaceae and *Neisseria*). The results remain largely unchanged (Figure S4). A gene arising in a lineage after the last common ancestor that has a similarity lower than that of the 99% of comparisons between the genes of the core genome is regarded as having diverged excessively to result from a duplication process and we infer that it arose by horizontal gene transfer.

2) Positional information. The remaining recently acquired genes are more similar to at least one member of the gene family than 1% of the pairwise similarities within families of the core genome. These may have arisen by gene duplication or by transfer from genomes containing closely related copies of the gene, e.g. conspecifics. Many experimental and in silico studies indicate that genetic duplications in prokaryotes create tandem repeats [8,38,39]. Since rearrangements occur at low rates in prokaryotes, recent paralogs are expected to be co-localized. Hence, we regard co-localized highly similar expansions as paralogs and non-colocalized expansions as xenologs. Using the syntenic map provided by the core genome and the core genome genes as markers of gene order in the ancestral genome, if two gene expansions are located in the same syntenic block and not separated by core genome genes, then they are regarded as co-localized. For paralogs or xenologs that fall in positions corresponding to synteny breakpoints in a genome, if they are inside the same breakpoint region they are considered to be co-localized. In any other situation we consider that the position is not conserved. Although, strictly speaking, genetic duplications occur in tandem, we used this more flexible definition of co-localization because pairs of paralogs can be separated by insertions and

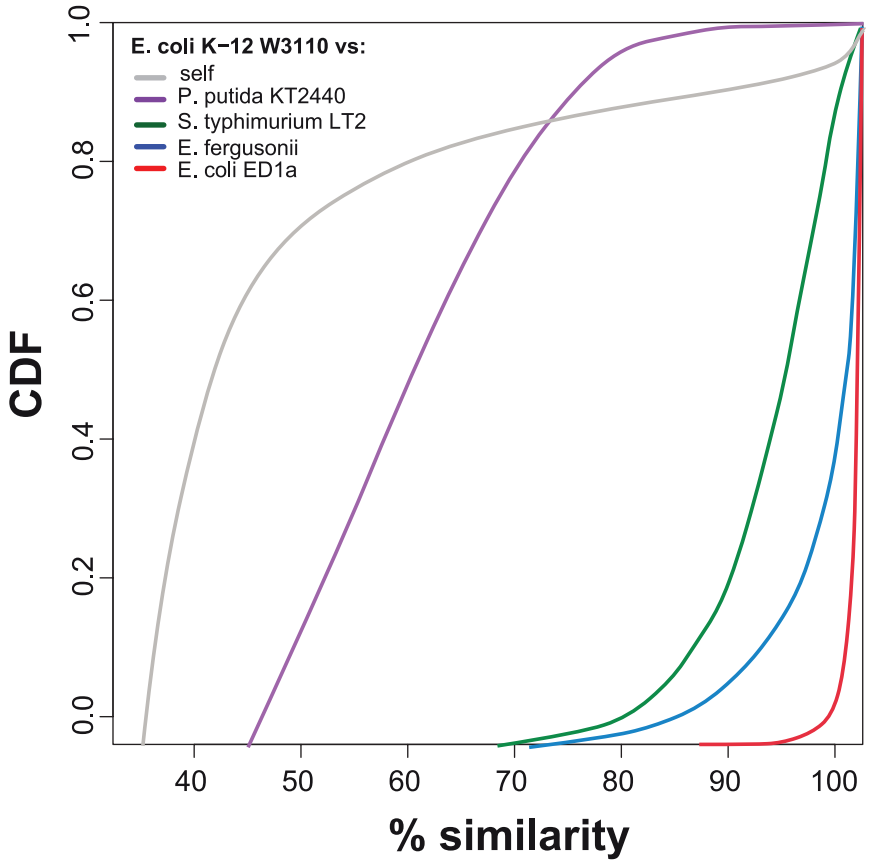


Figure 7. Cumulative distribution function plot of protein similarity. Colored lines correspond to CDF plots of the similarity between orthologous proteins of the core genome for the comparison of *E. coli* K12 W3110 with genomes of increasing phylogenetic distances. The gray line corresponds to the similarity between homologous genes in the *E. coli* K12 W3110 genome.
doi:10.1371/journal.pgen.1001284.g007

deletions after duplication (these newly arising spacing genes are not part of the core genome).

Supporting Information

Figure S1 M-GCAT visualization of core genome regions in 41 enterobacterial genomes. The core genome is represented as the vertical colored polygons common to all of the 41 genomes listed. Inversions can be identified as inverted rectangles. Syntenic regions of the core genome are colored according to the rainbow spectrum and meant to give a general idea of genome rearrangement. For example, violet colored regions generally remain in the same positions in all genomes, although there are some exceptions.
Found at: doi:10.1371/journal.pgen.1001284.s001 (1.12 MB PDF)

Figure S2 CDF plot of core genome similarity. CDF plot of percent similarity in the core genome and calculated thresholds for all clades except enterobacteria.
Found at: doi:10.1371/journal.pgen.1001284.s002 (0.61 MB PDF)

Figure S3 Phylogenetic trees for core genomes using BIONJ and Maximum Likelihood (ML) distances for all clades.
Found at: doi:10.1371/journal.pgen.1001284.s003 (1.52 MB PDF)

Figure S4 Effects of varying core genome similarity thresholds on HGT predictions.
Found at: doi:10.1371/journal.pgen.1001284.s004 (0.10 MB PDF)

Table S1 Clade summary. List of genomes, accession numbers, and summary.
Found at: doi:10.1371/journal.pgen.1001284.s005 (0.16 MB DOC)

Table S2 Phyletic patterns. Patterns of expansion in all 8 clades. Occurrences of synologs per genome. 0 = no gene present in genome, 1 = gene present w/o synologs, 2 = synologs in genome. Phyletic Pattern is the sequence of states for all 41 genomes for each of the families with synologs. Count indicates the frequency each phyletic pattern of synologs in the families.
Found at: doi:10.1371/journal.pgen.1001284.s006 (0.08 MB DOC)

Table S3 Interpretation of changes in ancestral states. The events marked with an asterisk can be interpreted as multiple successive events, e.g. 2 successive gene losses. The events in italics were ignored either because (a) there was no change from the ancestral node or (b) they cannot be parsimoniously explained by gene duplication.
Found at: doi:10.1371/journal.pgen.1001284.s007 (0.03 MB DOC)

Acknowledgments

The authors thank the Sanger Center for providing sequence data before publication. We also would like to thank Sara Vieira-Silva, Marie

Touchon, and Guillaume Achaz for critical reading of the manuscript. We are grateful to GA and Howard Ochman for fruitful discussions and suggestions. Finally, we would like to thank three anonymous reviewers for their constructive comments and criticisms.

Author Contributions

Conceived and designed the experiments: TJT EPCR. Performed the experiments: TJT. Analyzed the data: TJT EPCR. Contributed reagents/materials/analysis tools: TJT. Wrote the paper: TJT EPCR.

References

1. McCutcheon JP, McDonald BR, Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet 5: e1000565. doi:10.1371/journal.pgen.1000565.
2. Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, et al. (2007) Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol 25: 1281–1289.
3. Pasek S, Risler JL, Brezellec P (2006) The role of domain redundancy in genetic robustness against null mutations. J Mol Biol 362: 184–191.
4. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. Genome Biol 8: R51.
5. Wagner A (2008) Gene duplications, robustness and evolutionary innovations. Bioessays 30: 367–373.
6. Francino MP (2005) An adaptive radiation model for the origin of new gene functions. Nat Genet 37: 573–577.
7. Kugelberg E, Kofoid E, Reams AB, Andersson DI, Roth JR (2006) Multiple pathways of selected gene amplification during adaptive mutation. Proc Natl Acad Sci U S A 103: 17319–17324.
8. Andersson DI, Hughes D (2009) Gene amplification and adaptive evolution in bacteria. Annu Rev Genet 43: 167–195.
9. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938–950.
10. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, et al. (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zoolog B Mol Dev Evol 308: 58–73.
11. Demuth JP, Hahn MW (2009) The life and death of gene families. Bioessays 31: 29–39.
12. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97–108.
13. Alm E, Huang K, Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol 2: e143. doi:10.1371/journal.pcbi.0020143.
14. Serres MH, Kerr AR, McCormack TJ, Riley M (2009) Evolution by leaps: gene duplication in bacteria. Biol Direct 4: 46.
15. Treangen TJ, Abraham AL, Touchon M, Rocha EP (2009) Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev 33: 539–571.
16. Cho NH, Kim HR, Lee JH, Kim SY, Kim J, et al. (2007) The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. Proc Natl Acad Sci U S A 104: 7981–7986.
17. Goldman BS, Nierman WC, Kaiser D, Slater SC, Durkin AS, et al. (2006) Evolution of sensory complexity recorded in a myxobacterial genome. Proc Natl Acad Sci U S A 103: 15200–15205.
18. McLeod MP, Warren RL, Hsiao WW, Araki N, Myhre M, et al. (2006) The complete genome of Rhodococcus sp. RHA1 provides insights into a catabolic powerhouse. Proc Natl Acad Sci U S A 103: 15582–15587.
19. Lindroos H, Vinnere O, Mira A, Repsilber D, Naslund K, et al. (2006) Genome rearrangements, deletions, and amplifications in the natural population of Bartonella henselae. J Bacteriol 188: 7426–7439.
20. Evlampiev K, Isambert H (2008) Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc Natl Acad Sci U S A 105: 9863–9868.
21. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36: 492–496.
22. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.
23. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3: e130. doi:10.1371/journal.pbio.0030130.
24. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19: 2226–2238.
25. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11: 472–477.
26. Kunin V, Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. Genome Res 13: 1589–1594.
27. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res 16: 1099–1108.
28. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A 105: 10039–10044.
29. Snel B, Bork P, Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12: 17–25.
30. Hooper SD, Berg OG (2003) Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol 4: R48.
31. Gevers D, Vandepoele K, Simillon C, Van de Peer Y (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol 12: 148–154.
32. Pushker R, Mira A, Rodriguez-Valera F (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol 5: R27.
33. Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53: 673–684.
34. Wagner A (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Mol Biol Evol 23: 723–733.
35. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 24: 969–981.
36. van Passel MW, Marri PR, Ochman H (2008) The emergence and fate of horizontally acquired genes in Escherichia coli. PLoS Comput Biol 4: e1000059. doi:10.1371/journal.pcbi.1000059.
37. Rocha EP (2008) Evolutionary patterns in prokaryotic genomes. Curr Opin Microbiol 11: 454–460.
38. Romero D, Palacios R (1997) Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet 31: 91–111.
39. Achaz G, Rocha EPC, Netter P, Coissac E (2002) Origin and fate of repeats in bacteria. Nucleic Acids Res 30: 2987–2994.
40. Howell-Adams B, Seifert HS (2000) Molecular models accounting for the gene conversion reactions mediating gonococcal pilin antigenic variation. Mol Microbiol 37: 1146–1158.
41. Aras RA, Kang J, Tschumi AI, Harasaki Y, Blaser MJ (2003) Extensive repetitive DNA facilitates prokaryotic genome plasticity. Proc Natl Acad Sci U S A 100: 13579–13584.
42. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, et al. (2001) Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A 98: 15056–15061.
43. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, et al. (2001) Recombination within natural populations of pathogenic bacteria: short- term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98: 182–187.
44. Sharp PM, Li WH (1987) The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–1295.
45. Masuda T, Saito N, Tomita M, Ishihama Y (2009) Unbiased quantitation of Escherichia coli membrane proteome using phase transfer surfactants. Mol Cell Proteomics 8: 2770–2777.
46. Parmley JL, Hurst LD (2007) How common are intragene windows with KA>KS owing to purifying selection on synonymous mutations? J Mol Evol 64: 646–655.
47. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383–397.
48. Vernikos GS, Thomson NR, Parkhill J (2007) Genetic flux over time in the Salmonella lineage. Genome Biol 8: R100.
49. Rocha EPC, Danchin A (2004) An analysis of determinants of protein substitution rates in Bacteria. Mol Biol Evol 21: 108–116.
50. Hu P, Janga SC, Babu M, Diaz-Mejia JJ, Butland G, et al. (2009) Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol 7: e96. doi:10.1371/journal.pbio.1000096.
51. Martinez-Nunez MA, Perez-Rueda E, Gutierrez-Rios RM, Merino E (2010) New insights into the regulatory networks of paralogous genes in bacteria. Microbiology 156: 14–22.
52. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol 9: R4.
53. Lercher MJ, Pal C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25: 559–567.
54. Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. Proc Natl Acad Sci U S A 106: 21748–21753.
55. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318: 1449–1452.
56. Vernikos G, Thomson N, Parkhill J (2007) Genetic flux over time in the Salmonella lineage. Genome Biology 8: R100.
57. Isambert H, Stein RR (2009) On the need for widespread horizontal gene transfers under genome size constraint. Biol Direct 4: 28.
58. Rocha EPC (2006) Inference and Analysis of the Relative Stability of Bacterial Chromosomes. Mol Biol Evol 23: 513–522.
59. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.

60. Ochman H, Liu R, Rocha EP (2007) Erosion of interaction networks in reduced and degraded genomes. J Exp Zoolog B Mol Dev Evol 308: 97–103.
61. Wellner A, Lurie MN, Gophna U (2007) Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol 8: R156.
62. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9: 605–618.
63. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc Natl Acad Sci U S A 103: 11647–11652.
64. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456: 239–244.
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
66. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
67. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34: D32–36.
68. Fouts DE (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 34: 5839–5851.
69. Treangen TJ, Darling AE, Achaz G, Ragan MA, Messeguer X, et al. (2009) A novel heuristic for local multiple alignment of interspersed DNA repeats. IEEE/ACM Trans Comput Biol BioInf 6: 180–189.
70. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.
71. Rocha EP, Touchon M, Feil EJ (2006) Similar compositional biases are caused by very different mutational effects. Genome Res 16: 1537–1547.
72. Azad RK, Lawrence JG (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic Acids Res 35: 4629–4639.
73. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
74. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
75. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685–695.
76. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13: 555–556.

Emms and Kelly *Genome Biology* (2019) 20:238
https://doi.org/10.1186/s13059-019-1832-y

Genome Biology

SOFTWARE Open Access

OrthoFinder: phylogenetic orthology inference for comparative genomics

David M. Emms and Steven Kelly*

Abstract

Here, we present a major advance of the OrthoFinder method. This extends OrthoFinder's high accuracy orthogroup inference to provide phylogenetic inference of orthologs, rooted gene trees, gene duplication events, the rooted species tree, and comparative genomics statistics. Each output is benchmarked on appropriate real or simulated datasets, and where comparable methods exist, OrthoFinder is equivalent to or outperforms these methods. Furthermore, OrthoFinder is the most accurate ortholog inference method on the Quest for Orthologs benchmark test. Finally, OrthoFinder's comprehensive phylogenetic analysis is achieved with equivalent speed and scalability to the fastest, score-based heuristic methods. OrthoFinder is available at https://github.com/davidemms/OrthoFinder.

Keywords: Ortholog inference, Gene tree inference, Gene duplication, Comparative genomics

Background

Determining the phylogenetic relationships between gene sequences is fundamental to comparative biological research. It provides the framework for understanding the evolution and diversity of life on Earth and enables the extrapolation of biological knowledge between organisms. Given the central importance of this process to multiple areas of biological research, a diverse array of software tools have been developed that attempt to identify these relationships given sets of user-supplied gene sequences [1–3]. The majority of these software tools try to deduce phylogenetic relationships between gene sequences through heuristic analyses of pairwise sequence similarity scores (or expectation values) obtained from an all-vs-all BLAST [4] search, or accelerated alternatives to BLAST such as DIAMOND [5] or MMseqs2 [6]. Widely used methods include InParanoid [7], OrthoMCL [8], OMA [9], and OrthoFinder [10] all of which take different approaches to interrogating sequence similarity scores, and all of which produce different outputs—some identify orthogroups, some identify orthologs and paralogs, and some do both. As they each adopt different approaches to analyzing sequence similarity scores, each of the methods exhibits different performance characteristics on commonly used benchmark databases [1, 11].

Heuristic analysis of pairwise sequence similarity scores has historically been used to estimate the phylogenetic relationship between genes as it is readily computationally tractable. The central premise underlying their use is that higher scoring sequence pairs are likely to have diverged more recently than lower scoring sequence pairs. Thus, heuristic analysis of sets of pairwise sequence similarity scores can be used to estimate the phylogenetic relationships between sets of genes [7–9, 12, 13]. However, such score-based estimates of the phylogenetic relationship between genes are confounded by multiple factors. For example, variable sequence evolution rates between genes frequently lead to both false-positive and false-negative errors [14, 15] (Fig. 1). Such errors can be mitigated by the analysis of phylogenetic trees of genes [17], as phylogenetic trees are able to distinguish variable sequence evolution rates (branch lengths) from the order in which sequences diverged (tree topology) and hence clarify orthology and paralogy relationships (Fig. 1).

A number of tree-based online databases of orthologs have been developed including PhylomeDB [18], Ensembl-Compara [19], EggNOG [20], and TreeFam [21]. These highly used resources provide the user with the ability to explore the evolutionary history of genes using phylogenetic trees, giving a more complete picture than just pairwise orthology and paralogy relationships alone. Comparative analyses of these methods using standard benchmarking approaches have found no significant difference in ortholog

* Correspondence: steven.kelly@plants.ox.ac.uk
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

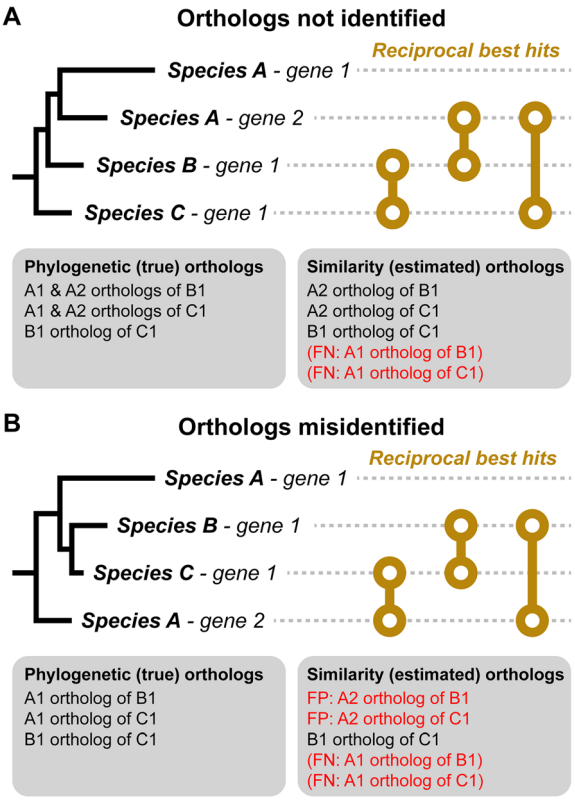


Fig. 1 Pairwise similarity score-based ortholog inference can be misled by variable sequence evolution rates. **a** Phylogenetic tree of a typical gene family. The correct orthologs of species A-gene 1 are not identified. **b** Species A-gene 2 is misidentified as the ortholog of the genes from species B and C. Left-hand side: gene trees with branches to scale and the true orthology relationships, which can be determined from the gene tree. Right-hand side: Reciprocal best hits (RBH) based on gene similarity scores that are monotonic with branch length and the orthology relationships inferred from these scores using standard heuristics (orthologs inferred using RBHs and co-orthology identified from within species hits better than closest RBH [8, 16]). FP, false positive; FN, false negative

detection accuracy of these online databases and score-based software tools [1], suggesting that the advantages of a phylogenetic approach have not yet been fully realized. Moreover, the pipelines and methodologies behind these online databases are generally not provided for users to run their own analyses. Thus, there is a need for an automated software tool that effectively exploits the phylogenetic approach to increase accuracy, but with the ease of use, speed, and scalability of a score-based heuristic method.

While an automated software tool for phylogenetic orthology inference from gene sequences is an important goal, the implementation of such a method presents several technical challenges. These comprise the following: (1) inferring a complete set of gene trees for all genes of a given set of species in a time-scale that is competitive with score-based heuristic methods; (2) automatically rooting these gene trees so that they can be correctly interpreted [22] without requiring the user to know the rooted species tree in advance; and (3) interpreting the gene trees to identify gene duplication events, orthologs, and paralogs while being robust to processes such as gene duplication, loss, incomplete lineage sorting, and gene tree inaccuracies. If these challenges could be addressed in a resource and time-efficient manner, then such a phylogenetic method would provide a step change for orthology inference, enabling the transition from similarity score-based estimates of phylogenetic relationships to phylogenetically delineated phylogenetic relationships between genes.

Some of the challenges listed above have been addressed in isolation by a range of bioinformatic methods. For example, there are a range of methods for identifying orthogroups of genes from user-supplied gene sequences [8–10, 12, 23] and a wide variety of gene tree inference methods that can infer trees from these orthogroups [24–28]. Similarly, there is a range of methods for inferring orthologs from gene trees that also vary in terms of scalability and accuracy [29–32]. However, other critical challenges had no existing solutions. For example, the inference of a complete set of rooted gene trees from a set of species proteomes would be a complex, multi-step process and generally require prior knowledge of the species tree. Equally, methods to infer orthologs from gene trees did not exist that were robust to processes such as incomplete lineage sorting and gene tree inference error while also being scalable to the large-scale analysis required for whole-genome orthology inference across hundreds of species. Thus, substantial technical challenges needed to be addressed to enable fully automated, accurate, and efficient phylogenetic delineation of the phylogenetic relationships between genes.

Here, we present a major update to OrthoFinder that addresses these challenges and significantly extends the scope of the original method. The updated version of OrthoFinder identifies orthogroups as in the original implementation [10] but then uses these orthogroups to infer gene trees for all orthogroups and analyzes these gene trees to identify the rooted species tree. The method subsequently identifies all gene duplication events in the complete set of gene trees and analyzes this information in the context of the species tree to provide both gene tree and species tree-level analysis of gene duplication events. Finally, the method analyzes all of this phylogenetic information to identify the complete set of orthologs between all species and provide extensive comparative genomics statistics. The complete OrthoFinder phylogenetic orthology inference method is accurate, fast, scalable, and customizable and is performed with a single command using only protein sequences as input.

Results

OrthoFinder algorithm overview and summary of results files

The OrthoFinder algorithm is described in detail in the "Methods" section. In brief, it addresses the challenges identified above in five major steps: (a) orthogroup inference, (b) inference of gene trees for each orthogroup, (c and d) analysis of these gene trees to infer the rooted species tree, (e) rooting of the gene trees using the rooted species tree, and (f–h) duplication-loss-coalescence (DLC) analysis of the rooted gene trees to identify orthologs and gene duplication events (mapped to their locations in both the species and gene trees) (Fig. 2). Thus, starting from just gene sequences, OrthoFinder infers orthogroups, orthologs, the complete set of gene trees for all orthogroups, the rooted species tree, and all gene duplication events and computes comparative genomic statistics. To illustrate the standard outputs provided by an OrthoFinder analysis, a graphical example of the complete set of results produced by OrthoFinder for ten metazoan species is shown in Fig. 3a–h.

The default, and fastest, version of OrthoFinder uses DIAMOND [24] for sequence similarity searches. These sequence similarity scores provide both the raw data for orthogroup inference [10] and for gene tree inference of these orthogroups using DendroBLAST [24]. The default implementation of OrthoFinder has been designed to enable a complete analysis with maximum speed and scalability using only gene sequences as input. However, OrthoFinder has also been designed to allow the use of alternative methods for tree inference and sequence search to accommodate user preferences. For example, BLAST [4] can be used for sequence similarity searches in place of DIAMOND. Similarly, gene trees do not need to be inferred using DendroBLAST. Instead, OrthoFinder can automatically infer multiple sequence alignments and phylogenetic trees using most user-preferred multiple sequence alignment and tree inference methods. Moreover, if the species tree is known prior to the analysis, this can also be provided as input, rather than inferred by OrthoFinder. Thus, while OrthoFinder is designed to require minimal inputs and computation, it can be tailored to suit the computational and data resources available to the user.

OrthoFinder has the highest ortholog inference accuracy

The accuracy of key component algorithms of OrthoFinder has been independently assessed in this work and in dedicated publications [5, 10, 22, 24, 33]. To demonstrate the accuracy of the overall method, the orthologs identified by OrthoFinder using its default options, along with multiple different configurations, were submitted to the community-supported *Quest for Orthologs* benchmarking server for the 2011_04 dataset [1] (see the "Methods" section for details of the tests). This dataset had benchmark results for the largest number of other methods and so allowed the most comprehensive comparison to competitor methods. The results of all of these analyses are shown in Fig. 4a–l and supported by additional analyses in Additional file 1: Figure S1-S3 and Additional file 1: Table S1.

The SwissTree and TreeFam-A tests within *Quest for Orthologs* assess the accuracy of ortholog inference against orthologs from gold-standard trees. For these tests, precision, recall, and *F*-score can be calculated. On these tests, the default, fastest version of OrthoFinder was 3–24% (SwissTree, Fig. 4a) and 2–30% (TreeFam-A, Fig. 4b) more accurate than any other method. The other versions of OrthoFinder were a further 1–3% more accurate than default OrthoFinder. No method was consistently second best to OrthoFinder.

For the *Quest for Orthologs* Standard and Generalized Species Tree Discordance Tests (STDT and GSTDT), no ground truth orthologs are known, and the methods are assessed on the percentage of trials in which a set of orthologs is identified across a set of species and the Robinson-Foulds distance between species tree and the gene tree of the putative orthologs. As such, standard precision, recall, and *F*-score measures cannot be calculated. For these tests, a "pseudo-*F*-score" was calculated using the percentage of the recovered ortholog sets in place of recall and 1 - normalized Robinson-Foulds distance in place of precision (equivalently, the proportion of bipartitions in an agreement between the species tree and the putative orthologs tree). On both STDT and GSTDT, all versions of OrthoFinder had an equal or higher pseudo-*F*-score than all versions of all other methods. The default, fastest version of OrthoFinder was 0–45% (STDT, Fig. 4c) and 10–59% (GSTDT, Fig. 4d) higher scoring than competing methods. The other versions of OrthoFinder were a further 1–6% higher scoring than the default version.

All versions of OrthoFinder, irrespective of algorithmic options, inferred more orthologs (higher recall/recovered ortholog sets) than any other tested method at a similar level of precision (Fig. 4e–l). Across the four tests, the default and fastest version of OrthoFinder (DIAMOND) achieved between 0 (Fig. 4g) and 65% (Fig. 4h) higher recall/recovered ortholog sets than competing methods. It achieved precision/ortholog species tree agreement between 5% lower (Fig. 4h) and 15% higher (Fig. 4g) than competing methods. Similarly, on the latest, 2018, benchmarks, all three versions of OrthoFinder were more accurate than all other methods on all four of the benchmarks: STDT, GSTDT, SwissTree, and TreeFam-A (Additional file 1: Figure S2).

In addition to testing OrthoFinder against competitor methods that can be run on raw sequence data,

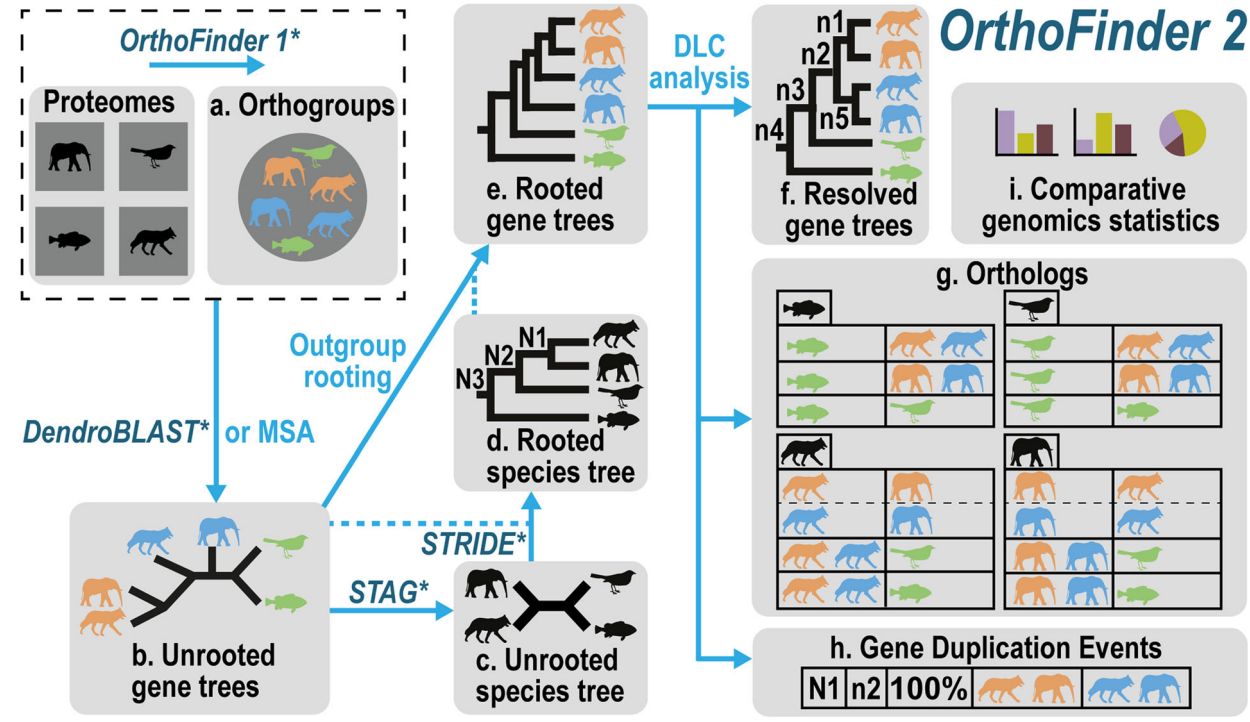


Fig. 2 The OrthoFinder workflow. The method used for each step is shown by the arrow. Published algorithms are shown in italics and are followed by an asterisk. A dotted blue line connecting with a solid arrow indicates additional data that are used in order to carry out the transformation indicated by the solid arrow. MSA, multiple sequence alignment-based tree inference; DLC, duplication-loss-coalescence. (**a**) Orthogroup inference using the original OrthoFinder algorithm (an orthogroup is the set of genes descended from a single gene in the last common ancestor of all the species under consideration). (**b**) Gene tree inference. (**c**) Species tree inference. (**d**) Species tree rooting (**e**) Gene tree rooting (**f**) Hybrid overlap + DLC analysis of rooted gene trees to infer orthologs and gene duplication events. (**g**) Illustration of the ortholog results table for the genes in each input species (four main boxes). The horizontal divisions within these show the orthologs for each individual species pair. (**h**) Illustration of the gene duplication event table showing the location of the gene duplication events mapped to the species tree, the location in the gene tree, the percent retention of the duplicate genes in the sampled species, and the genes descended from the gene duplication event. (**i**) Comparative genomics statistics

OrthoFinder was also compared with static database methods that involve various levels of human curation. All versions of OrthoFinder, irrespective of algorithmic options, had a higher *F*-score/pseudo-*F*-score across each of the four tests than any of the databases with the one exception of PANTHER on the Species Tree Discordance Test (Additional file 1: Figure S3). Overall, the default version *F*-score/pseudo-*F*-score was between 2 and 14% higher than the database methods. OrthoFinder (BLAST + MSA) scored between 5 and 17% higher than the database methods (Additional file 1: Figure S3). Thus, although OrthoFinder is fully automated and requires no manual curation, it also achieved higher accuracy than curated online database methods.

OrthoFinder is fast and scales well to hundreds of species

To demonstrate the scalability of the OrthoFinder method, it was run on sets of between 4 and 256 fungal species with 16 parallel processes (Fig. 4m). All other publicly available software tools that have been benchmarked on the *Quest for Orthologs* dataset were similarly tested. The default version of OrthoFinder ran in 192 s on the 4 species and 1.8 days on the 256 species datasets. In this time, it inferred orthogroups, all gene trees, the rooted species tree, orthologs, and gene duplication events (Fig. 4n). Overall, OrthoFinder was the second quickest method, with the fastest method SonicParanoid taking 1.2 days on the same 256 species set. Both OrthoFinder and SonicParanoid scaled well to the largest datasets, both taking less than half the time of the next best method (4.1 days, Fig. 4m).

There was a large range of runtimes across the complete set of methods. Many methods were unsuited to larger species sets, with 64 species being the largest set on which all methods were runnable within the 120 h (5 days) cutoff. At this point of comparison, the slowest method took 200 times longer to run than OrthoFinder. It should also be noted that no competitor method also provides gene trees or identifies gene duplication events (Fig. 4n). Thus, not only is OrthoFinder the most accurate method and the

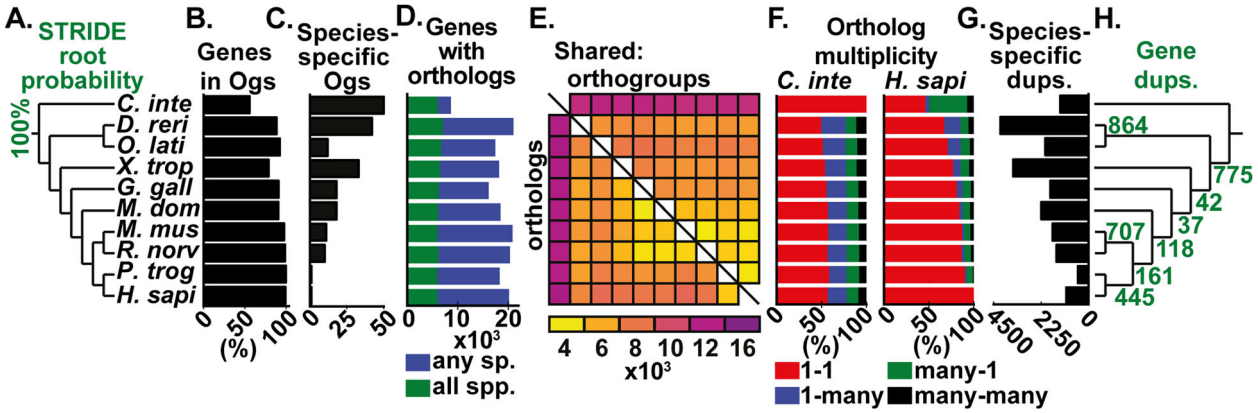


Fig. 3 Summary of OrthoFinder analysis of a set of Chordata species: *Ciona intestinalis*, *Danio rerio*, *Oryzias latipes*, *Xenopus tropicalis*, *Gallus gallus*, *Monodelphis domestica*, *Mus musculus*, *Rattus norvegicus*, *Pan troglodytes*, and *Homo sapiens*. Bar charts and heat map contain data for each species, aligned to the corresponding species in the tree in **a**. **a** The species tree inferred by STAG and rooted by STRIDE. **b** Percentage of genes from each species assigned to orthogroups. **c** The number of species-specific orthogroups. **d** The number of genes with orthologs in any/all species. **e** Heat map of the number of orthogroups containing each species pair (top right) and orthologs between each species (bottom left). **f** Ortholog multiplicities for two species, *C. intestinalis* and *H. sapiens*, with respect to all other species. **g** The number of gene duplication events on each terminal branch of the species tree. **h** The number of duplications on each branch of the species tree and retained in all descendant species. OG, orthogroup; sp., species; spp., species (plural); dups., gene duplication events

second fastest method, it also provides the largest quantity of phylogenomic information.

OrthoFinder efficiently and accurately solves the challenge of inferring a rooted species tree from unaligned protein sequence data

Rooted gene trees are required to enable the use of phylogenetic information for ortholog inference, since the correct placement of the root is required for the correct dissection of phylogenetic relationships between genes in the tree [22]. However, the vast majority of tree inference methods infer unrooted trees. Gene trees can be correctly rooted given the knowledge of the underlying rooted species tree, and thus, OrthoFinder first infers and then roots the species tree for the set of species being analyzed. OrthoFinder solves these two challenges (species tree inference and rooting) using two algorithms developed specifically for this purpose.

The species tree is inferred from the set of unrooted orthogroup gene trees using STAG [33], and this species tree is rooted using STRIDE [22]. STAG was developed to leverage the vast amount of phylogenetic information already available in the complete set of orthogroup gene trees inferred by OrthoFinder. It was also developed to be robust to high levels of gene duplication and loss that can hamper methods that rely on sets of single-copy orthologs [33]. It outperformed popular species tree inference methods on benchmark data and scaled well to large datasets [33].

Methods for *ab initio* species tree rooting (i.e., without prior knowledge of a suitable outgroup) have received little attention [22]. STRIDE was similarly developed to leverage gene duplication events in the complete set of orthogroup gene trees to efficiently determine the root of the species tree and achieved high accuracy on benchmark data [22]. The ability of OrthoFinder to automatically leverage the raw amino acid sequence data to infer the rooted species tree thus enables outgroup rooting of the complete set of orthogroup gene trees for any input set of species and for all gene trees. This is a critical step for enabling phylogenetic orthology inference from gene sequences.

OrthoFinder implements a novel duplication-loss-coalescent algorithm for identifying gene duplication events and orthologs

Given a set of rooted orthogroup gene trees, the final major challenge in accurately dissecting phylogenetic relationships between genes is to account for incomplete lineage sorting and gene tree error. Existing methods for determining if genes within a gene tree are orthologs or paralogs either had poor accuracy or were unable to scale to the number and size of the orthogroup gene trees that must be analyzed. Thus, to address this challenge, a novel, scalable algorithm based on the duplication-loss-coalescent model was developed (see the "Methods" section).

To demonstrate the relative performance characteristics of this method, it was applied to two independent simulated datasets [32, 34] and compared to three popular, comparable methods: GSDI Forester [29], DLCpar (full and search) [32], and species overlap method [31] (Fig. 5). It was also compared to Notung [30], but since branch support values were not available, which Notung

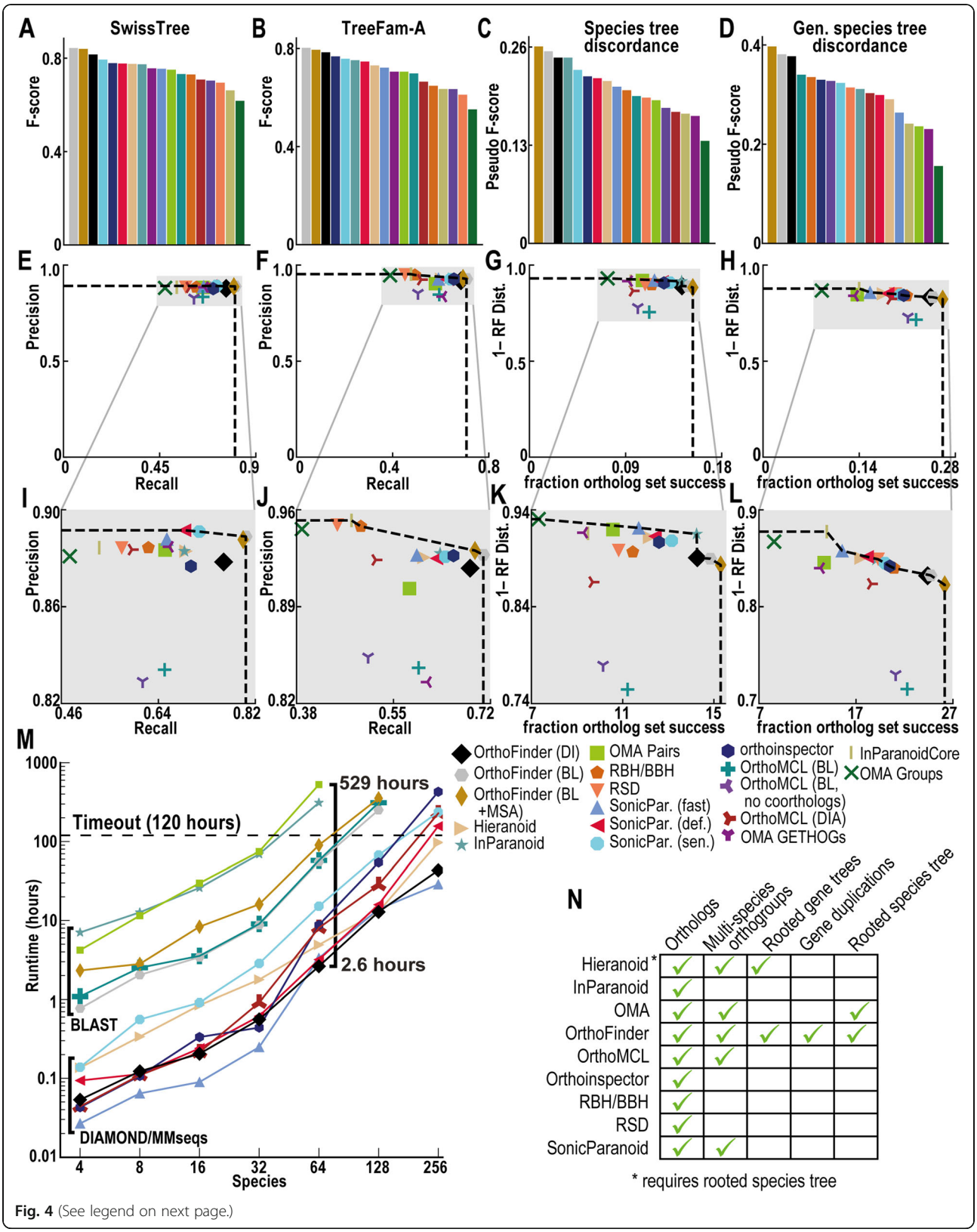


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 a–l Quest for Orthologs 2011_04 benchmarks (see [1]) on 66 species across Eukarya, Bacteria and Archaea for ortholog inference methods. Dotted line shows Pareto frontier. Data for graphs are in Additional file 1: Table S1. **a**, **b** *F*-score on SwissTree and TreeFam-A tests. **c** "Pseudo-*F*-score" across the two Species Tree Discordance Tests (STDT). **d** "Pseudo-*F*-score" across the four Generalized Species Tree Discordance Tests (GSTDT). **e**–**f** Agreement of orthologs SwissTree/FreeFam-A trees **g**-**h** Benchmarks across the STDT & GSTDT. X-axis: Total fraction of randomly selected genes with predicted orthologs in a predefined set of species for the two STDTs & four GSTDTs respectively. Y-axis: Average (1 – normalised Robinson-Foulds distance) between gene tree for putative orthologs and the known species tree across the two STDT & four GSTD respectively. The four individual GSTDTs and two individual STDTs are shown in Additional file: 1 **i-l** Zoom in of plots **e**-**h**. See Methods section "Ortholog Benchmarking" for details of Quest for Orthologs benchmarks. **m** Runtime for each method with 4-256 input Fungi proteomes. **n** Results returned by methods, a multi-species orthogroup is the set of genes descended from a single gene in the last common ancestor of three or more species

use to improve its accuracy, it achieved identical results to Forester and so is not shown as an additional method here. In terms of accuracy, the novel OrthoFinder method outperformed all methods other than DLCpar (full) (Fig. 5a, Additional file 1: Table S2). However, DLCpar (full) was unable to analyze realistic-sized species datasets. For example, while the OrthoFinder method was able to analyze the complete set of 18,651 orthogroup gene trees (948,449 genes) from 128 fungal species in 141 s, DLCpar (full) was unable to process a considerably smaller, 4-species dataset (2259 trees, 12, 958 genes) in 120 h (Fig. 5b). Thus, OrthoFinder is the most accurate method that is scalable to realistic datasets. This algorithm enables accurate interrogation of orthogroup gene trees in a manner that can analyze thousands of gene trees across hundreds of species in minutes on standard computing hardware (Fig. 5b).

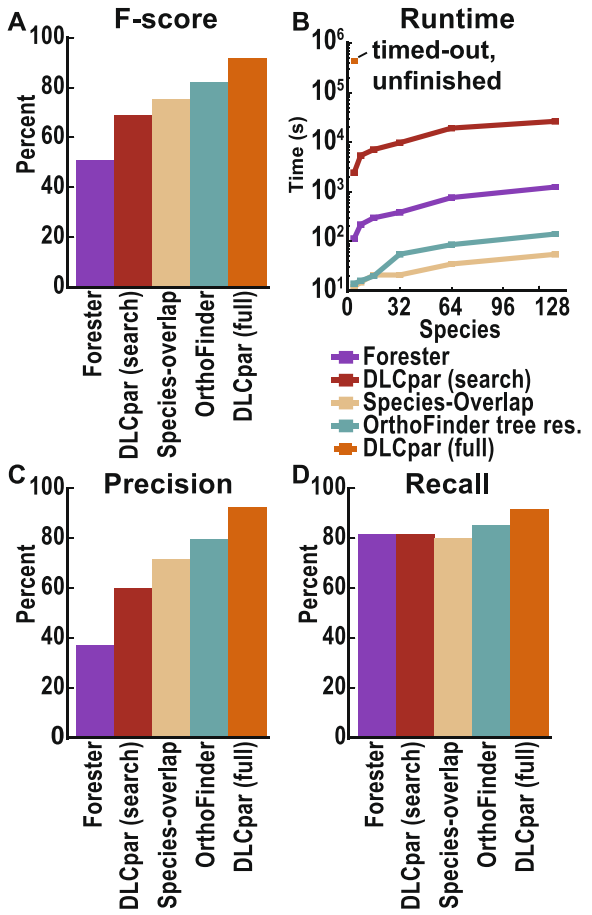


Fig. 5 a Duplication F-score, on simulated gene trees. **b** Runtime to analyse all trees from the 4 to 128 species Fungi datasets (see Methods), a maximum time of 120 hours (4.3x10 seconds) was allowed. DLCpar (full) did not complete the smallest dataset in this time limit and so only the lower bound for the first time point is shown. **c-d** Precision and recall

Discussion and conclusions

Phylogenetic relationships between gene sequences are defined by their relationship in a gene tree in the context of a species tree. Due to the complexity of conducting phylogenetic orthology inference from raw gene sequences, multiple methods have been developed to bypass phylogeny and approximate phylogenetic relationships from heuristics on pairwise sequence similarity scores. Such approximations are subject to common errors that are avoidable by the analysis of phylogenetic trees of gene sequences. Here, we present a substantial update to OrthoFinder that provides an easy-to-use, fast, accurate, and fully phylogenetic orthology inference software tool.

From testing on community standard benchmarks, we demonstrate that OrthoFinder is the most accurate orthology inference method available. Furthermore, we show that by taking a phylogenetic approach, OrthoFinder provides substantial additional information (including rooted gene trees, rooted species trees, and gene duplication events) that are not provided by heuristic methods. Thus, OrthoFinder is the most accurate and most data-rich orthology inference method for comparative genomics.

The only input required for OrthoFinder is the set of amino acid sequences of the protein-coding genes for the species of interest. OrthoFinder has been designed with ease of use in mind, and the entire analysis is launched with a single command. The default parameters for OrthoFinder are optimized for speed and scalability and enable the combined analysis of hundreds of species on commonly available computer resources. However, OrthoFinder is also designed with the expert user in mind, and intermediate steps in the algorithm can be substituted with other methods for multiple

sequence alignment and tree inference should the user wish. We illustrate the time-accuracy trade-off associated with changes in the internal steps of the algorithm and show that the fastest and least accurate implementation OrthoFinder is still more accurate than any other orthology inference method.

Methods

OrthoFinder workflow

A gene tree is the canonical representation of the evolutionary relationships between the genes in a gene family. Thus, ortholog inference from gene trees is an important goal. However, no automated software tools are available that provide genome-wide ortholog inference from gene trees. A number of challenges had to be addressed to enable this. These included the efficient partitioning of genes into small, non-overlapping sets such that all orthologs of a gene are contained in the same set as the original gene; scalable and accurate inference of gene trees from these gene sets; automatic rooting of these gene trees without a user-provided species tree; and robust ortholog inference in the presence of imperfect gene tree inference. The OrthoFinder workflow was designed to address each of these challenges and is described in detail below.

By default, OrthoFinder infers orthologs from the orthogroup trees (a gene tree for the orthogroup) using the steps shown in Fig. 2. Input proteomes are provided by the user using one FASTA file per species. Each file contains the amino acid sequences for the proteins in that species. Orthogroups are inferred using the original OrthoFinder algorithm [10]; an unrooted gene tree is inferred for each orthogroup using DendroBLAST [24]; the unrooted species tree is inferred from this set of unrooted orthogroup trees using the STAG algorithm [33]; this STAG species tree is then rooted using the STRIDE algorithm by identifying high-confidence gene duplication events in the complete set of unrooted orthogroup trees [22]; the rooted species tree is used to root the orthogroup trees; orthologs and gene duplication events are inferred from the rooted orthogroup trees by a novel hybrid algorithm that combines the "species-overlap" method [31] and the duplication-loss-coalescent model [32] (described below); and comparative statistics are calculated. All major steps of the algorithm are parallelized to allow optimal use of computational resources. Only the orthogroup inference was provided in the original implementation of OrthoFinder [10]; all other subsequent steps are new and described below.

Use of orthogroups for gene tree inference

Orthologs are the set of genes in a species pair descended from a single gene in the last common ancestor of those two species. An orthogroup is the set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of that set of species. Thus, an orthogroup is the natural extension of orthology to multiple species.

For ortholog inference, orthogroups are the optimum partitioning of genes for gene tree inference: An orthogroup is the smallest set of genes such that, for all genes it contains, the orthologs of these genes are also in the same set. Since gene tree inference scales super-linearly with the number of genes, partitioning genes into the smallest possible sets is the most efficient way of constructing a set of gene trees that encompass all orthology relationships. Although partitioning genes into larger sets (e.g., gene families containing gene duplication events prior to the LCA) would decrease the number of gene trees to be inferred, the super-linear scaling of gene tree inference would result in a longer overall runtime for the complete set of trees. The original OrthoFinder orthogroup inference method is still the most accurate method on the independent Orthobench test set [10] and thus is used for this step.

Customizable steps in the OrthoFinder method

There are two customizable steps in the OrthoFinder method: (1) the sequence search method and (2) the orthogroup tree inference method. The default option for step 1 is DIAMOND [5]. The default option for step 2 is DendroBLAST [24]. The default options are recommended by the authors as they are fast and achieve high accuracy on the Quest for Orthologs benchmarks [1] (Fig. 4a–d). However, the user is free to substitute any alternative methods for these steps. Currently, supported methods for step 1 include BLAST [4] and MMseqs2 [6]. Similarly, any combination of multiple sequence alignment and tree inference method can be substituted in for step 2. For illustrative purposes, the default multiple sequence alignment method is MAFFT [35] and the default tree inference method is FastTree [25]; this combination is benchmarked above. It is impossible for the authors to test all possible combinations of multiple sequence alignment and tree inference methods, and the selected methods were chosen because of their speed and scalability characteristics [25, 35]. OrthoFinder provides flexibility for the user to select their preferred method. More accurate multiple sequence alignment and tree inference methods should give more accurate ortholog inference, and many studies exist comparing the accuracy and runtime characteristics of the available methods [36, 37]. A user-editable configuration file is provided in JSON format that allows new sequence search, multiple sequence alignment, and tree inference methods to be added to OrthoFinder. To facilitate the trialing of alternative multiple sequence alignment and tree inference methods, OrthoFinder provides the option to restart an existing

analysis after the orthogroup inference stage. This skips the requirement to compute the all-vs-all sequence search and orthogroup inference and thus accelerates testing of different internal steps.

Species tree inference and rooting

The rooted species tree is required in order to identify the correct out-group in each orthogroup tree, as correct gene tree rooting is critical for the orthology assessment from that tree [22]. Since orthogroups can potentially contain any subset of the species in the analysis, it is not sufficient to simply know the out-group for the complete species set. Instead, the complete rooted species tree is required. If the user knows the rooted species tree for the set of species being analyzed, then it is recommended to specify this tree manually at the command line to remove the possibility of species tree inference error. Such a tree can be provided as a Newick format text file. In the event that a species tree is not provided (or not known), then OrthoFinder automatically infers it.

Sets of one-to-one orthologs that are present in all species are often used for species tree inference; however, in real-world large-scale analyses, these can be rare [33]. A new algorithm, Species Tree from All Genes (STAG), was developed to allow species tree inference even for species sets with few or no complete sets of one-to-one orthologs present in all species [33]. Without this algorithm, species tree inference could fail if there were no sets of one-to-one orthologs present in all species. STAG infers the species tree using the most closely related genes within single-copy or multi-copy orthogroups. In benchmark tests, STAG [24] had higher accuracy than other leading methods for species tree inference, including maximum likelihood species tree inference from concatenated alignments of protein sequences, ASTRAL [38] and NJst [39].

The Species Tree Root Inference from Duplication Events (STRIDE) algorithm [22] is used to root the species tree in OrthoFinder. STRIDE was developed to enable the rooting of the species tree using only information available in the set of gene trees. STRIDE does this by identifying the set of well-supported in-group gene duplication events in the complete set of unrooted orthogroup trees, and using these events to infer a probability distribution over an unrooted STAG species tree for the location of its root. Similarly to STAG, STRIDE has been shown to identify the correct root of the species tree in multiple large-scale molecular phylogenetic data sets spanning a wide range of time scales and taxonomic groups [22]. In some cases, it is possible that there could be few duplications within the gene trees, and so STRIDE will not be able to identify the root of the species tree, or will only be able to exclude the root from clades in which gene duplication events are observed. In this case, ortholog inference should still not be significantly impacted since the rooting of the gene tree only affects ortholog inference in cases where gene duplication events are present [22]. This makes the STRIDE approach particularly suited to gene tree rooting for ortholog inference.

Gene tree rooting

Tree inference methods infer unrooted gene trees. A gene tree must be correctly rooted in order for it to show the correct evolutionary history of the gene family and thus to allow correct ortholog inference. The orthogroup trees could contain any subset of the input species. In general, the rooted species tree, inferred as described above, can be used to root the orthogroup trees by identifying the out-group clade in each orthogroup tree and placing the root on the branch separating this out-group from the remaining genes.

However, species tree and gene tree topologies can arise in which this simple approach will not work, and so, a robust generalization of this outgroup rooting method is required in order to be able to root any potential gene tree. Firstly, in the species tree, the out-group could consist of a single species or multiple species. Secondly, in the gene tree, the genes from the out-group could be in a monophyletic clade or there may be no bipartition in the tree that separates all the genes from the out-group from all remaining genes. Thirdly, a gene duplication event could have occurred in the gene tree prior to the divergence of the out-group from the remaining species. Thus, the most ancient bipartition of the gene tree would be a gene duplication event separating the genes into two clades rather than a bipartition separating the out-group from the in-group. Such a gene tree should be rooted on this bipartition. Both of these two descendant clades could then potentially contain genes from both the out-group and in-group species. Thus, there will be no bipartition in such a tree that separates the genes of the out-group species from the genes of the in-group species.

The algorithm used by OrthoFinder searches for the correct bipartition on which to place the root. For each bipartition in the gene tree, it calculates two scores. The first, S_{AD}, quantifies how well the bipartition corresponds to an ancient duplication prior to the divergence of the species. The second, S_{IO}, quantifies how well the bipartition corresponds to the divergence of the out-group species from the in-group species. Both S_{IO} and S_{AD} range between 0 and 1. Let O be the set of species in the out-group and I be the set of species in the in-group. For a bipartition in the unrooted gene tree, let A be the set of species with genes on one side of the bipartition and let B be the set of species with genes on the other side of the bipartition. Then:

$$S_{IO} = \frac{|O \cap A|}{|O|}\frac{|I \cap B|}{|I|}\left(1 - \frac{|O \cap B|}{|O|}\right)\left(1 - \frac{|I \cap A|}{|I|}\right),$$

$$S_{AD} = \frac{|\,O \cap A\,|}{|\,O\,|}\frac{|\,I \cap B\,|}{|\,I\,|}\frac{|\,O \cap B\,|}{|\,O\,|}\frac{|\,I \cap A\,|}{|\,I\,|}.$$

Each of the four terms in these equations quantifies the proportion of in-/out-group species the bipartition correctly includes/excludes from clade A/B of the gene tree (giving the $2^3 = 8$ terms in total across the two equations). The bipartition with the highest score for either S_{IO} or S_{AD} is the optimal root for the gene tree using this measure.

The effectiveness of these scores at identifying the correct root can be seen by considering the following. A bipartition with a value of 1 for S_{IO} implies that it perfectly divides the tree into an in-group and out-group and implies a value of 0 for S_{AD} for all bipartitions in the tree (thus, there are no potential bipartitions corresponding to an ancient duplication). This is the correct bipartition on which to root the tree since it separates the in-group from the out-group genes. Conversely, a bipartition with a value of 1 for S_{AD} implies that the bipartition is a duplication event before the divergence of any of the species, with all species present for both duplicates. It implies a value of 0 for S_{IO} for all bipartitions in the tree (thus, there is no bipartition that corresponds to a first speciation event that splits the genes into an out-group clade and an in-group clade). The highest value for either S_{IO} or S_{AD} across the tree shows that the corresponding bipartition is close to one of these perfect cases and is the best root for the gene tree.

Ortholog inference and identification of gene duplication events from gene trees

A number of methods were considered for distinguishing orthologs from paralogs in gene trees. Duplication and loss reconciliation, e.g., Forester, uses a rooted species tree and rooted gene tree to determine if each node in the gene tree is a speciation or a duplication event. Genes that diverged at a speciation event are orthologs whereas those that diverged at a duplication event are paralogs. DLCpar [32] uses a model for duplication-loss-(deep) coalescent (DLC) that addresses incongruence between the gene and species trees to increase accuracy. It exists in two versions which we label DLCpar (full) and DLCpar (search). DLCpar (full) considers the complete space of possible reconciliations to find the maximum parsimony solution under the DLC model but can have large runtimes even for relatively small gene trees. DLCpar (search) instead employs an iterative search for a locally optimal solution, which can differ from the globally optimal solution. A third approach, here referred to as the species-overlap method, is employed in a number of ortholog databases [20, 31] and was originally described in a method for determining orthologs of human genes [31]. In this method, nodes in the gene tree are identified as duplication nodes if the sets of species below its child nodes overlap; otherwise, the node is a speciation node. Genes that diverged at a speciation node are orthologs, and those that diverged at a duplication node are paralogs.

These methods were tested on the fungal orthogroups (in parallel, using 16 cores) to determine their runtime on sets of typical orthogroup trees derived from sets of between 4 and 128 species. Our implementation of the species-overlap method was the fastest, taking 55 s to analyze the largest datatset (Fig. 5). This dataset consisted of the 18,651 orthogroup trees containing 948,449 genes and corresponded to the complete set of orthogroup trees for the 128 fungal species. Forester was 21 times slower, and DLCpar (search) was over 500 times slower. DLCpar (full) was unable to complete the analysis of the smallest input dataset in 120 h and so was not tested on any of the larger datasets. To put this time in context, all steps in the OrthoFinder algorithm for this dataset collectively take less than 4 min in total (i.e., orthogroup inference, gene tree inference, species tree inference, species tree rooting, gene tree rooting).

To compare the accuracy of the above methods, they were each tested for their precision and recall in identifying gene duplication events on simulated "flies" and "primates" datasets [32] and a simulated "metazoa" dataset [34]. Since for all methods tested a node in a gene tree is either a duplication or speciation event, the identification of all gene duplication events is equivalent (by complementation) to the identification of all speciation events. Thus, the overall accuracy at identifying gene duplication events is equivalent to the overall accuracy at identifying orthologs. The most accurate method on the simulated data was DLCpar (full) with an *F*-score of 91.8% followed by the species-overlap method with an *F*-score of 75.5%.

Since DLCpar (full) was the most accurate method on the simulated datasets but was unsuitable for analyzing gene trees with more than four species a novel hybrid algorithm was developed. This aimed to combine the strengths of the highest accuracy DLCpar (full) method with simplifications from the species-overlap method to achieve high accuracy in a reasonable runtime.

In the DLC model, clades of genes containing no duplicates are analyzed to find the most parsimonious reconciliation with the species tree. This is required since the goal for DLCpar is a complete reconciliation of the gene tree with the species tree. However, in the species-overlap method, clades of single-copy genes are identified as orthologs without further analysis of the topology of their relationship. This assumption is reasonable,

since trees of single-copy orthologs are frequently topologically distinct from the species tree. For example, in an analysis of 1030 gene trees of one-to-one orthologs from 23 fungi species, all 1030 gene trees were topologically distinct from each other and from the species tree [40]. The analysis of such clades under the DLC model is likely to be computationally costly with no benefit in terms of accuracy of ortholog inference.

On the other hand, when a gene duplication event has occurred, it is important to accurately identify the genes affected by this event since the location of the event determines which genes are orthologs and which are paralogs. In the hybrid algorithm developed for OrthoFinder, these nodes, for which there is evidence of a gene duplication event through overlapping species sets, are analyzed under the DLC model. The DLC model is used to attempt to find the most parsimonious interpretation of this node in terms of which genes diverged at the gene duplication event and which diverged at a speciation event.

As described, this method would still require exploring a large search space for the nodes under consideration, and the reduction in runtime would not be significant. Thus, to accelerate the process, duplication and loss events are inferred directly using the species-overlap method. A duplication event is inferred from an overlap in the species sets below a node and a loss event is inferred by the presence of a gene from a species in one of the descendant clades but not in the other. The analysis can then be accelerated by classifying a node according the species overlaps of its subclades up to a maximum total topological depth of two below the node being analyzed (clades O, Additional file 1: Figure S4A). The possible sub-cases for the overlaps between these clades have been enumerated (Additional file 1: Figure S4B). For each sub-case, the most parsimonious interpretation under the DLC model has been pre-calculated (Additional file 1: Figure S4C) and can thus be corrected without the need for a topology search.

The algorithm implemented in OrthoFinder is as follows. A post-order traversal of the orthogroup tree is performed (a node is not visited until all its descendant nodes have been visited), analyzing each node of the orthogroup tree in turn. A given node is analyzed to identify if the species sets below its child nodes overlap. If there is an overlap, the smallest sub-clade below each child node that contains the complete set of overlapping species is identified up to a maximum total topological depth of two below the node (clades O, Additional file 1: Figure S4A). The node is assigned to the corresponding sub-case (Additional file 1: Figure S4B). If a more parsimonious interpretation of the sub-case is available under the DLC model, then the sub-tree below the node is rearranged to match this interpretation (Additional file 1: Figure S4C). After the node has been analyzed, the next node in the post-order traversal is analyzed. Note, the choice of a post-order traversal allows the traversal to be continued unimpeded despite any such rearrangements below the node being analyzed. The resulting gene trees are referred to as "resolved" gene trees and correspond to the "locus tree" under the DLCpar model [32]. Orthologs and gene duplication events are determined from the resolved gene tree according to the species overlap method.

Although only a single traversal of the tree is employed, rather than the iterative search and rearrangement employed by DLCpar, the post-order traversal enables more parsimonious interpretations of child clades below a node to be identified prior to the analysis of the parent node. Thus, the analysis of sub-trees below a node informs the subsequent analysis of the node itself. In theory, nodes could be categorized to sub-cases based on the overlaps of clades at a greater topological depth than that employed here. This conservative approach was taken since the number of subcases increases exponentially, and a total topological depth of two proved sufficient to achieve a higher accuracy for the method compared to the simple species overlap. The analysis of clades to this depth proved sufficient to increase the *F*-score from 72% with just the species-overlap method to 80% with the hybrid algorithm (Fig. 5a). The pre-calculated solutions for each sub-case removed the need for costly, iterative search using random (i.e., unguided) tree rearrangement operations thus accelerating the analysis considerably. The hybrid algorithm was able to analyze the complete set of orthogroup trees for the 128 fungi species in 141 s; this was 9 times faster than Forester and 187 times faster than DLCpar (search) (Fig. 5d). The hybrid method also outperformed both methods in terms of accuracy (Fig. 5a). Note that the species tree is not required for the hybrid model used by OrthoFinder. The only use of the species tree is in determining the root for each orthogroup tree. All gene tree processing is performed using the python ETE toolkit [41].

Simulation tests of OrthoFinder gene duplication event inference accuracy

The tests for gene duplication event inference accuracy were performed on the simulated "flies" and "primates" dataset from [32] and a simulated "metazoa" dataset from [34]. To model real data, the flies and primate datasets used known species trees, parameters for divergence times, duplication rates, loss rates, population sizes, and generation times. Trees were simulated with varying effective population sizes and duplication rates so as to model incomplete lineage sorting [32, 34]. The flies dataset consisted of 12,000 trees with 12 species and 12,032 gene duplication events. The primates dataset

consisted of 7500 trees with 17 species and 16,066 gene duplication events. The metazoa dataset intended to emulate the complexity of real data by using heterogeneity in rates of duplication and loss, a complex model of sequence evolution, and then inferring trees with a homogenous, simple model [34]. It consisted of 2000 gene trees with 40 species and 4967 gene duplication events. For comparison, Forester [29], DLCpar (full), DLCpar (search) [32], and the overlap algorithm (i.e., without OrthoFinder's tree resolution) were also tested.

All methods were provided with the input rooted gene tree and, where appropriate, the rooted species tree (Forester and DLCpar). No other parameters required specification for any of the other methods. The rooted gene trees were provided as part of the simulated data for the flies and primates datasets. Multiple sequence alignment (MSA) files were provided for the metazoa dataset. For this dataset, gene trees were inferred from the MSAs using FastTree so as to also include a potential level of tree inference error and were rooted with reconroot [32]. The OrthoFinder rooting algorithm was not used so as to avoid inadvertently biasing the results in favor of OrthoFinder. All methods were provided with the same input rooted gene trees. The complete set of gene duplication events identified by each of the methods was compared against the ground truth gene duplication events. An inferred gene duplication was identified as correct if the two sets of genes observed post-duplication exactly matched the two sets of genes post-duplication from the ground truth data.

The performance testing of the methods for identifying gene duplication events was performed on the orthogroup trees from the 4- to 128-species Fungi datasets as inferred by OrthoFinder with default parameters. The commands for Forester and DLCpar were run in parallel using GNU Parallel [42] using 16 threads on these gene trees. The OrthoFinder method was run via the "scripts/resolve.py" program included as part of the OrthoFinder distribution. To allow testing, the species-overlap method was also implemented in OrthoFinder and was run using the same program with the option "--no_resolve."

Ortholog benchmarking

Orthogroup inference accuracy of OrthoFinder has already been tested using the independent Orthobench dataset [11]. This showed to be the most accurate method tested in terms of overall *F*-score (although other methods scored higher in terms of either precision or recall while scoring proportionally worse in the other) [10]. The community developed "Quest for Orthologs" benchmarks [1] were used to assess the accuracy of the newly developed OrthoFinder ortholog inference using the 2011_04 dataset. This dataset had benchmarks for the largest set of methods and so provided the widest comparison with other methods. OrthoFinder was tested using the default method (DIAMOND sequence search and DendroBLAST trees, no additional options). It was also tested with the BLAST replacing DIAMOND (options: "-S blast") and with both BLAST search and multiple sequence alignment and maximum likelihood tree inference (options: "-S blast -M msa"). In the latter, MAFFT [35] and FastTree [25] were used for multiple sequence alignment and tree inference as described above. For each of these three cases, OrthoFinder was run on the 66 reference proteomes of the Quest for Orthologs test set with a single command ("-f Proteomes/" + options), and the inferred orthologs were submitted to the Quest for Orthologs web server for benchmarking.

The Quest for Orthologs benchmarks are described in detail in [1]. The Species Tree Discordance Test and the generalized version of this test both consider a set of species partitioned into clades with a known species tree topology connecting the clades. The benchmarking consists of a repeated test. For one of the clades of species, a gene is selected at random for each instance of the test. If the orthology inference method under scrutiny predicts an ortholog for that gene for at least one species from each of the remaining clades, then the test is recorded as a "successful ortholog set." For each successful ortholog set, an MSA is constructed and a gene tree inferred using RAxML [28]. The normalized Robinson-Foulds (RF) distance is calculated between this tree and the known species tree. The result of the benchmark is the fraction of successful ortholog sets and the average RF distance for these successful sets. A higher fraction of success and a lower average RF distance indicates a better ortholog inference method under this test. The benchmarks include two different Species Tree Discordance Tests (STDT) across two different species sets and four Generalized Species Tree Discordance Tests (GSTDT) across four different species sets. In Fig. 4g, h, the total fraction of successful ortholog sets and the average normalized RF distance across these successful ortholog sets across the two/four species sets are reported for the STDT and GSTDT. The individual GSTDT and STDT results for the four individual species sets are given in Additional file 1: Figure S1.

Minor changes have been made to the labeling and orientation of the axes compared to the presentation in the Quest for Orthologs paper [1] to improve the consistency with the SwissTree and TreeFam-A benchmarks. The altered *y*-axis for the GSTDT and STDT presented here is (1 - normalized RF distance) so that higher *y* values always correspond to the better agreement with the species tree for all benchmark figures. The number of completed ortholog set successes for the STDT and GSTDT is reported as a fraction rather than the total number. For the SwissTree and TreeFam-A tests, the axes are labeled as "precision" instead of "pos. predictive value rate" and

"recall" instead of "true positive rate" as this is more standard terminology for the quantities reported by the tests.

The full set of benchmarks, the input files, and the ortholog inference results can be seen online at http://orthology.benchmarkservice.org/. A comprehensive summary of the benchmarks, as described above, is show in Fig. 4a–l for ortholog prediction software tools. The corresponding comparisons against online databases are shown in Additional file 1: Figure S2 and Additional file 1: S3. The complete datasets are available to download from Zenodo research archive at https://doi.org/10.5281/zenodo.1481147 [43].

Performance testing

We constructed sets of fungal proteomes of increasing size for performance testing. Ensembl Genomes was interrogated on 6 November 2017 using its REST API [44] to identify all available fungal genomes. To achieve an even sampling of species, we selected 1 species per genera and excluded genomes from candidate phyla or phyla with fewer than 3 sequenced genomes. This gave a set of 272 species which were downloaded from the Ensembl FTP site [45]. We created datasets of increasing size by randomly selecting 4, 8, 16, 32, 64, 128, and 256 species such that the last common ancestor was the same for each dataset. Each dataset was analyzed using a single Intel E5-2640v3 Haswell node (16 cores) on the Oxford University ARCUS-B server using 16 parallel threads for OrthoFinder with DIAMOND (arguments: "-S diamond -t 16 -a 16"). The complete datasets for all analyzed species subsets are available for download from Zenodo at https://doi.org/10.5281/zenodo.1481147. All methods submitted to Quest for Orthologs that provided a user-runnable implementation of the method were tested on the same fungi datasets and the same ARCUS-B server nodes and run in parallel using 16 threads (when supported by the method).

Chordata dataset

The data for the OrthoFinder analysis of the ten Chordata species for the illustration of the results of an OrthoFinder analysis (Fig. 2a–h) are provided in the Zenodo archive https://doi.org/10.5281/zenodo.1481147. This includes the input proteomes, the OrthoFinder results, and the script used to generate the figures from the results. OrthoFinder was run with default settings (DIAMOND sequence search and DendroBLAST gene trees).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1832-y.

Additional file 1. Supplementary figures and tables.

Additional file 2. Review history.

Acknowledgements

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work.

Review history

The review history is available as Additional file 2.

Additional information

Peer review information: Barbara Cheifet and Tim Sands were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

DE and SK conceived and designed the project. DE developed the algorithms. DE and SK discussed the results and wrote the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. SK is a Royal Society University Research Fellow.

Availability of data and materials

The OrthoFinder source code and executables are available at https://github.com/davidemms/OrthoFinder [46] and are released under the GNU General Public License, GPL-3.0 license. The 256 Fungi proteomes and the Chordata proteomes datasets were prepared for this study and are available in the Zenodo research data archive at https://doi.org/10.5281/zenodo.1481147 [42]. The reference proteomes for the Quest for Orthologs benchmarks are available from ftp://ftp.ebi.ac.uk/pub/databases/reference_proteomes/previous_releases/. The orthologs inferred by all methods and all benchmark results are available from the Quest for Orthologs Benchmark Server [1]: https://orthology.benchmarkservice.org. The metazoa dataset [34] and flies and primates [32] datasets used for the genome duplication accuracy analysis are available from the respective authors.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 24 April 2019 Accepted: 23 September 2019
Published online: 14 November 2019

References

1. Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, et al. Standardized benchmarking in the quest for orthologs. Nature Methods. 2016;13:425.
2. Nichio BTL, Marchaukoski JN, Raittz RT. New tools in orthology analysis: a brief review of promising perspectives. Front Genet. 2017;8:165.
3. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for gene orthology inference. Brief Bioinform. 2011;12:379–91.
4. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
5. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
6. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
7. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010;38:D196–203.
8. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

9. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 2011;39:D289–94.
10. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
11. Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. Orthology prediction methods: a quality assessment using curated protein families. Bioessays. 2011;33:769–80.
12. Cosentino S, Iwasaki W. SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics. 2019;35:149–51.
13. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: comprehensive orthology analysis and visual exploration. Bmc Bioinformatics. 2011;12:11.
14. Lafond M, Miardan MM, Sankoff D. Accurate prediction of orthologs in the presence of divergence after duplication. Bioinformatics. 2018;34:366–75.
15. Fitch WM. Distinguishing homologous from analogous proteins. Sys Zool. 1970;19:99.
16. Remm M, Storm CEV, Sonnhammer ELL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001;314: 1041–52.
17. Dalquen DA, Dessimoz C. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol. 2013; 5:1800–6.
18. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldon T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014;42:D897–902.
19. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. Ensembl comparative genomics resources. Database. 2016;2016:baw053. https://academic.oup.com/database/article/doi/10.1093/database/baw053/2630361.
20. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldon T, Rattei T, Creevey C, Kuhn M, et al. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 2014;42: D231–9.
21. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li RQ, Liu T, Zhang Z, Bolund L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34:D572–80.
22. Emms DM, Kelly S. STRIDE: species tree root inference from gene duplication events. Mol Biol Evol. 2017;34(12):3267–78.
23. Schreiber F, Sonnhammer ELL. Hieranoid: hierarchical orthology inference. J Mol Biol. 2013;425:2072–81.
24. Kelly S, Maini PK. DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. PLoS One. 2013;8:e58537.
25. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
26. Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol. 2015;32: 2798–800.
27. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
28. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
29. Zmasek CM, Eddy SR. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics. 2001;17:821–8.
30. Chen K, Durand D, Farach-Colton M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 2000;7:429–47.
31. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. The human phylome. Genome Biol. 2007;8(6):R109.
32. Wu YC, Rasmussen MD, Bansal MS, Kellis M. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Res. 2014;24:475–86.
33. Emms D, Kelly S. STAG: species tree inference from all genes. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/267914v1.
34. Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale coestimation of species and gene trees. Genome Res. 2013;23:323–30.
35. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
36. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One. 2011;6:e18093.
37. Zhou XF, Shen XX, Hittinger CT, Rokas A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Mol Biol Evol. 2018;35:486–503.
38. Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 2015; 31:44–52.
39. Liu L, Yu LL. Estimating species trees from unrooted gene trees. Syst Biol. 2011;60:661–7.
40. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013;497:327.
41. Huerta-Cepas J, Serra F, Bork P: ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution 2016, 33:1635–38.
42. Tange O. GNU Parallel - the command-line power tool. login. 2011;36:42–7.
43. Emms D, Kelly S: Supplemental dataset for: OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. https://doi.org/10.5281/zenodo.1481147 2019.
44. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, Ruffier M, Taylor K, Vullo A, Flicek P. The Ensembl REST API: Ensembl data for any language. Bioinformatics. 2015;31:143–5.
45. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43:D662–9.
46. Emms D, Kelly S: OrthoFinder. GitHub. https://github.com/davidemms/OrthoFinder. 2019. Accessed 21 Oct 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



A.4 Phylogeny

Recommended reading for Chapter 4 (listed in Section 4.7).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Saitou and Nei (1987) — neighbour joining; the distance method still used everywhere, and a model of a clear algorithm paper.
  DOI · via SU library (SU login)

  The neighbor-joining method: a new method for reconstructing phylogenetic trees. Saitou & Nei (1987). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040454

  > A new method called the neighbor-joining method is proposed for reconstructing phylogenetic trees from evolutionary distance data. The principle of this method is to find pairs of operational taxonomic units (OTUs [= neighbors]) that minimize the total branch length at each stage of clustering of OTUs starting with a starlike tree. The branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method. Using computer simulation, we studied the efficiency of this method in obtaining the correct unrooted tree in comparison with that of five other tree-making methods: the unweighted pair group method of analysis, Farris's method, Sattath and Tversky's method, Li's method, and Tateno et al.'s modified Farris method. The new, neighbor-joining method and Sattath and Tversky's method are shown to be generally better than the other methods.

- Felsenstein (1981) — maximum likelihood trees from DNA sequences and the pruning algorithm.
  DOI · via SU library (SU login)

  Evolutionary trees from DNA sequences: A maximum likelihood approach. Felsenstein (1981). *Journal of Molecular Evolution*. doi:10.1007/bf01734359

  > The application of maximum likelihood techniques to the estimation of evolutionary trees from nucleic acid sequence data is discussed. A computationally feasible method for finding such maximum likelihood estimates is developed, and a computer program is available. This method has advantages over the traditional parsimony algorithms, which can give misleading results if rates of evolution differ in different lineages. It also allows the testing of hypotheses about the constancy of evolutionary rates by likelihood ratio tests, and gives rough indication of the error of ;the estimate of the tree.

- Kapli et al. (2020) — tree building in the genomic age; an up-to-date review of methods, models and pitfalls.
  DOI · via SU library (SU login)

  Phylogenetic tree building in the genomic age. Kapli et al. (2020). *Nature Reviews Genetics*. doi:10.1038/s41576-020-0233-0

  > Knowing phylogenetic relationships among species is fundamental for many studies in biology. An accurate phylogenetic tree underpins our understanding of the major transitions in evolution, such as the emergence of new body plans or metabolism, and is key to inferring the origin of new genes, detecting molecular adaptation, understanding morphological character evolution and reconstructing demographic changes in recently diverged species. Although data are ever more plentiful and powerful analysis methods are available, there remain many challenges to reliable tree building. Here, we discuss the major steps of phylogenetic analysis, including identification of orthologous genes or proteins, multiple sequence alignment, and choice of substitution models and inference methodologies. Understanding the different sources of errors and the strategies to mitigate them is essential for assembling an accurate tree of life.

- Minh, Schmidt, et al. (2020) — IQ-TREE 2; what a modern maximum-likelihood program does (models, search, support).
  full text below · DOI · open access (CC BY 4.0)

  IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Minh et al. (2020). *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa015

> IQ-TREE (http://www.iqtree.org, last accessed February 6, 2020) is a user-friendly and widely used software package for phylogenetic inference using maximum likelihood. Since the release of version 1 in 2014, we have continuously expanded IQ-TREE to integrate a plethora of new models of sequence evolution and efficient computational approaches of phylogenetic inference to deal with genomic data. Here, we describe notable features of IQ-TREE version 2 and highlight the key advantages over other software.

- Philippe et al. (2011) — why more sequences are not enough; systematic error in phylogenomics.
  full text below · DOI · open access (CC BY 4.0)

  Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. Philippe et al. (2011). *PLoS Biology*. doi:10.1371/journal.pbio.1000602

- Hug et al. (2016) — a new view of the tree of life from thousands of genomes.
  full text below · DOI · open access (CC BY 4.0)

  A new view of the tree of life. Hug et al. (2016). *Nature Microbiology*. doi:10.1038/nmicrobiol.2016.48

  > The tree of life is one of the most important organizing principles in biology(1). Gene surveys suggest the existence of an enormous number of branches(2), but even an approximation of the full scale of the tree has remained elusive. Recent depictions of the tree of life have focused either on the nature of deep evolutionary relationships(3-5) or on the known, well-classified diversity of life with an emphasis on eukaryotes(6). These approaches overlook the dramatic change in our understanding of life's diversity resulting from genomic sampling of previously unexamined environments. New methods to generate genome sequences illuminate the identity of organisms and their metabolic capacities, placing them in community and ecosystem contexts(7,8). Here, we use new genomic data from over 1,000 uncultivated and little known organisms, together with published sequences, to infer a dramatically expanded version of the tree of life, with Bacteria, Archaea and Eukarya included. The depiction is both a global overview and a snapshot of the diversity within each major lineage. The results reveal the dominance of bacterial diversification and underline the importance of organisms lacking isolated representatives, with substantial evolution concentrated in a major radiation of such organisms. This tree highlights major lineages currently underrepresented in biogeochemical models and identifies radiations that are probably important for future evolutionary analyses.

The open-access papers are reproduced below under their original licences.

IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era

Bui Quang Minh ,*[,1,2] Heiko A. Schmidt,[3] Olga Chernomor,[3] Dominik Schrempf,[3,4] Michael D. Woodhams,[5] Arndt von Haeseler,[†,3,6] and Robert Lanfear[†,2]

[1]Research School of Computer Science, Australian National University, Canberra, ACT, Australia

[2]Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia

[3]Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna, Austria

[4]Department of Biological Physics, Eötvös Lórand University, Budapest, Hungary

[5]Discipline of Mathematics, University of Tasmania, Hobart, TAS, Australia

[6]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: m.bui@anu.edu.au.

Associate editor: Emma Teeling

Abstract

IQ-TREE (http://www.iqtree.org, last accessed February 6, 2020) is a user-friendly and widely used software package for phylogenetic inference using maximum likelihood. Since the release of version 1 in 2014, we have continuously expanded IQ-TREE to integrate a plethora of new models of sequence evolution and efficient computational approaches of phylogenetic inference to deal with genomic data. Here, we describe notable features of IQ-TREE version 2 and highlight the key advantages over other software.

Key words: **phylogenetics, phylogenomics, maximum likelihood, models of sequence evolution.**

Brief Communication

IQ-TREE is a widely used and open-source software package for phylogenetic inference using the maximum likelihood (ML) criterion. The high performance of IQ-TREE results from the efficient integration of novel phylogenetic methods that improve the three key steps in phylogenetic analysis: fast model selection via ModelFinder (Kalyaanamoorthy et al. 2017), an effective tree search algorithm (Nguyen et al. 2015), and a novel ultrafast bootstrap approximation (Minh et al. 2013; Hoang et al. 2018). Zhou et al. (2018) independently showed that the tree search algorithm in IQ-TREE exhibits good performance in terms of both computing times and likelihood maximization when compared with other popular ML phylogenetics software such as RAxML (Stamatakis 2014) and PhyML (Guindon et al. 2010). IQ-TREE also plays a vital role in the software ecosystem for biomedical research. For instance, it is an integral component of many popular open-source applications such as Galaxy (Afgan et al. 2018), Nextstrain (Hadfield et al. 2018), OrthoFinder (Emms and Kelly 2015), and QIIME 2 (Bolyen et al. 2019).

Since the release of IQ-TREE version 1.0 in 2014, we have continuously developed IQ-TREE to integrate a plethora of new evolutionary models and efficient methods for analyzing large phylogenomic data sets. Here, we present IQ-TREE version 2 and highlight the key new features and improvements. To demonstrate its performance and compare it with other software, we used a 2-GHz CPU server to analyze 2 large sequence alignments: a DNA alignment of 110 vertebrate species and 25,919 sites (Fong et al. 2012) which we call the DNA-data set, and an amino acid alignment of 76 metazoan species and 49,388 sites (Whelan et al. 2017) which we call the AA-data set.

Time-Reversible Models of Sequence Evolution

IQ-TREE 2 supports more than 200 time-reversible evolutionary models, including all standard substitution models for DNA, protein, codon, binary, and multistate-morphological data (Felsenstein 2004; Lemey et al. 2009). Rate heterogeneity across sites can be accommodated either by the discrete Γ distribution (Yang 1994b) with invariant sites (Gu et al. 1995) or by a distribution-free rate model (Kalyaanamoorthy et al. 2017). Site-specific rates can be estimated by the empirical Bayesian method via the –rate option or by ML (Mayrose et al. 2004) via the –mlrate option. These estimated site-specific rates can be useful for downstream analysis such as quantification of phylogenetic informativeness, signal, and noise (Dornburg et al. 2016). For single nucleotide polymorphism or morphological data, the absence of invariant sites can be accounted for by an ascertainment bias correction (Lewis 2001).

Moreover, IQ-TREE 2 offers a number of advanced models for phylogenomic data including partitioned models (Lanfear et al. 2012; Chernomor et al. 2016), mixture models (Le et al. 2008, 2012; Le and Gascuel 2010), posterior-mean

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

site frequency models (Wang et al. 2018), and heterotachy models (Crotty et al. 2019). For allele-frequency data, IQ-TREE 2 implements polymorphism-aware models (Schrempf et al. 2016, 2019). With partitioned models, one can specify either a partition file (as in IQ-TREE 1) or a directory of single-locus alignments (a new feature in IQ-TREE 2). In the latter case, IQ-TREE 2 will load and concatenate all alignments within the directory, eliminating the need for users to manually perform this step. In addition to implementing existing mixture models, IQ-TREE 2 goes beyond the mixture models employed in PhyML-mixtures (Le et al. 2008) and RAxML-NG software (Kozlov et al. 2019), by allowing for user-defined mixture models using the "MIX{model_1,. . .,model_k}" syntax. IQ-TREE 2 is also substantially faster than PhyML-mixtures. For example, optimization of the model parameters for the LG4X model on the AA-data set took 1.9 min in IQ-TREE 2, 3 min in RAxML-NG, and 17.6 min in PhyML-mixtures.

Because the aforementioned substitution models assume time reversibility, IQ-TREE 1 only enabled inference of *unrooted* trees (Felsenstein 1981). In IQ-TREE 2, we included non-time-reversible models (e.g., Norris 1997), meaning that IQ-TREE 2 enables inference of *rooted* trees.

Nonreversible Substitution Models

IQ-TREE 2 allows users to reconstruct rooted trees using nonreversible models, a feature not available in most ML packages due to numerical and computational expense. We substantially revised the IQ-TREE code to overcome these obstacles. First, due to nonreversibility of the rate matrix Q, the naïve computation of the transition probability matrix $P(t) = e^{Qt}$, where t is the branch length, is unstable due to complex eigenvalues of Q (Moler and Loan 1978). Eigen-decomposition and scaling-squaring techniques, for example, as provided in the Eigen3 library (Guennebaud et al. 2010) remove the numerical problems. IQ-TREE 2 employs eigen-decomposition to diagonalize Q into its (complex) eigenvalues, eigenvectors and inverse eigenvectors, which are used to compute $P(t)$. If Q is not diagonalizable, then IQ-TREE 2 switches to the scaling-squaring technique to compute $P(t)$. The eigen-decomposition is fast but sometimes unstable, whereas the scaling-squaring is slow but stable. Second, IQ-TREE 2 uses a rooted tree data structure and an adjusted pruning algorithm for computing likelihoods of nonreversible models on rooted trees (Boussau and Gouy 2006). Third, we adapted the hill climbing nearest neighbor interchange, part of the tree search heuristic in IQ-TREE to account for rooted trees. Fourth, we introduced a root search operation that moves the root to the neighboring branches (by default up to 2 branches away from the current root branch) and retains the rooting position with the highest likelihood. Users can increase this parameter (–root-dist option) to test for the position of the root across more or all branches.

Based on these improvements, we could efficiently implement 99 nonreversible DNA models known as Lie Markov models (Woodhams et al. 2015), the unrestricted model (UNREST) for DNA (Yang 1994a) and—for the first time—the general nonreversible model for amino acid sequences that we call NONREV. For the DNA-data set (Fong et al. 2012), tree search under the general time-reversible model took 2.8 min using 4 CPU cores, whereas the UNREST model took 10.7 min. For the AA-data set (Whelan et al. 2017) tree search under the protein general time-reversible model took 2.52 h, whereas the NONREV model took 5.8 h. The implementation of nonreversible models in IQ-TREE 2 opens new avenues of evolutionary research.

Fast Likelihood Mapping Analysis

IQ-TREE 2 provides a fast and parallel implementation of the quartet likelihood mapping (Strimmer and von Haeseler 1997) to visualize phylogenetic information in alignments or to study the relationships of taxon-groups in large data sets. To this end, IQ-TREE 2 evaluates the exact ML value of all relevant quartets. Application of the original implementation of quartet likelihood mapping in TREE-PUZZLE (Schmidt et al. 2002) (i.e., with 10,000 random quartets and exact ML quartet evaluation) to the DNA-data set took 282 min, whereas the improved implementation in IQ-TREE 2 required only 1 min using one CPU core and 21 s using four cores. Similarly, analysis of the AA-data set took 99.5 h in TREE-PUZZLE, 17 min in IQ-TREE 2 with one CPU core, and 4.5 min using four cores. Together with the extended repertoire of sequence evolution models, likelihood mapping facilitates a thorough investigation of much larger sequence alignments.

New Options for Tree Search

IQ-TREE 2 allows users to perform a constrained tree search (-g option), such that the resulting ML tree will respect a set of user-defined splits, which may also contain polytomies. This option is helpful to enforce the monophyly of certain groups. A tree test (see below) can be performed to ensure that the constrained tree is not significantly worse than the unconstrained tree.

IQ-TREE 2 also provides a fast tree search (-fast option) using an algorithm resembling that implemented in FastTree2 (Price et al. 2010). Here, IQ-TREE 2 computes two starting trees using Maximum Parsimony and Neighbor-Joining (Gascuel 1997), which are then optimized by the hill climbing nearest neighbor interchange moves. For our example DNA-data set, the default tree search took 27 min, whereas the fast IQ-TREE search and FastTree2 needed 82 and 85 s, respectively. For the AA-data set, the default tree search took 5.7 h, whereas the fast IQ-TREE search and the FastTree2 search took 13.9 and 4.3 min, respectively. The speed of FastTree2 was accomplished at the cost of producing substantially worse trees than RAxML, PhyML, and IQ-TREE (Zhou et al. 2018).

IQ-TREE 2 also provides a –runs option, which conducts multiple independent tree searches, summarizes the resulting trees, and reports the tree(s) with the highest log-likelihood. This option is recommended for difficult data sets, for

example, with many taxa and/or limited phylogenetic signal, to increase the probability that the true ML tree is found.

Systematically Accounting for Missing Data

Missing data—when some loci or sites are absent for some species—are almost unavoidable in phylogenomic data sets. In the presence of missing data, each species tree can be associated with a corresponding set of induced single-locus trees. For each locus, an induced tree is obtained from the species tree by removing species with unavailable sequences for that locus. Missing data can create phylogenetic terraces (Sanderson et al. 2011), where for two or more species trees the associated sets of induced single-locus trees are identical, leading to identical likelihoods under an edge-unlinked partitioned model.

IQ-TREE 2 employs the terraphast library (Biczok et al. 2018) to automatically report the inferred ML trees that reside on a terrace. If so, users are advised to gather more data or filter out gappy taxa/loci. Moreover, IQ-TREE 2 generalizes the terrace concept to partial terraces (Chernomor et al. 2015), where a subset of the induced single-locus trees is identical. IQ-TREE 2 exploits partial terraces to improve tree search under partitioned models and achieves up to 4.5- and 8-fold speedups compared with IQ-TREE 1 and RAxML, respectively (Chernomor et al. 2016).

Single-Locus Tree Inference

IQ-TREE 2 enables users to infer individual locus trees (-S option), which can be used for subsequent coalescent (e.g., Mirarab et al. 2014) or concordance analyses (e.g., Minh et al. 2018). Users need to specify either a partition file that delineates the borders between loci or a directory containing the individual locus alignments in individual files. In both cases, IQ-TREE 2 will perform separate model selection and tree searches for each locus, which are automatically scheduled on the k CPU cores. The loci are ranked according to their expected computational costs, where the costs are estimated as a product of the number of sequences, distinct site patterns, and character states (4 for DNA and 20 for protein). The k loci with highest costs will be assigned to one of the k cores. When a core has finished computation the next locus in the ranked list will be assigned to that core. This process continues until all computations are complete.

We compared our scheduling approach with the program ParGenes version 1.0.1 (Morel et al. 2019), which uses RAxML-NG for tree search and a more sophisticated scheduling algorithm. For the DNA-data set, IQ-TREE 2 inferred 168 locus trees in 18.6 and 9.2 min using 2 and 4 CPU cores, respectively; whereas the same analysis in ParGenes took 9.17 and 5.98 min. For the AA-data set with 127 loci, IQ-TREE 2 took 2.05 and 1.05 h using 2 and 4 cores, respectively; whereas ParGenes needed 1.53 and 0.81 h. Therefore, IQ-TREE 2 shows a higher parallel efficiency (DNA: 101% and AA: 97.6%) (Grama et al. 2003) than ParGenes (DNA: 76.7% and AA: 94.4%), but ParGenes needs less computational time. The upshot of this is that both programs are likely to perform similarly on very large data sets.

Fast Branch Tests

IQ-TREE 2 provides fast and parallel implementations for several existing branch tests including the approximate likelihood ratio test (aLRT) (Anisimova and Gascuel 2006), the Shimodaira–Hasegawa-like aLRT (SH-aLRT) (Guindon et al. 2010), and the aBayes test (Anisimova et al. 2011). The SH-aLRT is parallelized over the bootstrap samples to maximize load balance and efficiency. These tests can be performed on the reconstructed ML tree or a user-defined tree. For the DNA-data set, PhyML version 3.3.20190321 (Guindon et al. 2010) took 33.9 h to perform the SH-aLRT, whereas IQ-TREE 2 needed only 1.5 min (1,300-fold speedup). On the AA-data set, PhyML took 173.5 h to perform the SH-aLRT tests, whereas IQ-TREE 2 needed just 1.6 min (a greater than 2,000-fold speedup).

Fast Topology Tests

IQ-TREE 2 also provides fast and parallel implementations of existing tree topology tests including the Shimodaira–Hasegawa test (Shimodaira and Hasegawa 1999), the approximately unbiased test (Shimodaira 2002), and the expected likelihood weight (Strimmer and Rambaut 2002). Moreover, IQ-TREE 2 is well suited for partitioned models because it provides the site, gene, and gene-site bootstrap resampling schemes (Hoang et al. 2018).

For the DNA-data set CONSEL (Shimodaira and Hasegawa 2001), the original and unparallelized implementation of the approximately unbiased test, took 5 min to test 100 tree topologies, whereas IQ-TREE 2 took 1.8 min with one CPU core and 1.1 min with four CPU cores. For the AA-data set, CONSEL took 9.4 min and IQ-TREE 2 took 8.7 min with one CPU core and 2.4 min with four CPU cores. IQ-TREE provides added convenience by calculating site log-likelihoods itself, rather than relying on site log-likelihood output provided by other software.

Scalability with Large Data Sets

IQ-TREE 2 implements several features to facilitate the analysis of large data sets. It uses multithreading to speed up computations in a range of areas and can automatically determine the best number of threads for the computer at hand. IQ-TREE 2 also parallelizes the computation across cluster nodes using Message Passing Interface (Snir et al. 1998). IQ-TREE 2 periodically writes a compressed checkpoint file that enables resumption of an interrupted analysis. IQ-TREE 2 also provides a memory-saving mode (Izquierdo-Carrasco et al. 2012), that is automatically invoked when the memory requirement exceeds the RAM size, and a safe mode (-safe option) to avoid numerical underflow for taxon-rich alignments (automatically invoked for data sets with $>$2,000 sequences).

We benchmarked the memory-saving mode, where the RAM consumption is reduced by half (-mem 0.5). For the DNA-data set, IQ-TREE 2 took 27 and 32.8 m (21% increase) under the full- and half-memory mode, respectively, whereas the same analysis needed 5.7 and 6.2 h (9% increase) for the AA-data set. This increase in computing times is insignificant

compared with 50% saving in memory, which could otherwise be a bottleneck for very large data sets and mixture models requiring hundreds of GB of RAM.

Documentation, User Support, and Workshop Materials

An extensive user manual, quick start guide, tutorials, and command reference are available (http://www.iqtree.org/doc, last accessed February 6, 2020). We actively maintain a forum (https://groups.google.com/d/forum/iqtree, last accessed February 6, 2020) for user support, bug reports, and feature requests. We regularly teach IQ-TREE at the Workshop on Molecular Evolution (https://molevolworkshop.github.io, last accessed February 6, 2020), the Workshop on Virus Evolution and Molecular Epidemiology (https://rega.kuleuven.be/cev/veme-workshop/, last accessed February 6, 2020), and the Workshop on Phylogenomics (http://evomics.org, last accessed February 6, 2020). Most workshop materials are freely available at http://www.iqtree.org/workshop/ (last accessed February 6, 2020).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Suha Naser, Cassius Manuel, and Lam-Tung Nguyen for fruitful discussions, Lukasz Reszczynski for integrating the terraphast library, Ian Brennan for the logo design, Dimitri Hoehler for checking the IQ-TREE code quality (https://github.com/adrianzap/softwipe/wiki/Code-Quality-Benchmark, last accessed February 6, 2020), Alexandros Stamatakis and an anonymous reviewer for constructive comments on the manuscript, and more than 100 users for helpful feedback and bug reports. Their names are listed in the full release notes (http://www.iqtree.org/release, last accessed February 6, 2020). This work was supported by the Austrian Science Fund (Grant No. I-2805-B29) to A.v.H., by the Australian National University Futures Scheme grant to R.L., and by the European Research Council under the European Unionã s Horizon 2020 research and innovation programme (Grant No. 714774) to D.S.

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46(W1):W537–W544.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol.* 55(4):539–552.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol.* 60(5):685–699.

Biczok R, Bozsoky P, Eisenmann P, Ernst J, Ribizel T, Scholz F, Trefzer A, Weber F, Hamann M, Stamatakis A. 2018. Two C plus plus libraries for counting trees on a phylogenetic terrace. *Bioinformatics* 34:3399–3401.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* 37(8):852–857.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol.* 55(5):756–768.

Chernomor O, Minh BQ, von Haeseler A. 2015. Consequences of common topological rearrangements for partition trees in phylogenomic inference. *J Comput Biol.* 22(12):1129–1142.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol.* 65(6):997–1008.

Crotty SM, Minh BQ, Bean NG, Holland BR, Tuke J, Jermiin LS, von Haeseler A. Forthcoming 2019. GHOST: recovering historical signal from heterotachously-evolved sequence alignments. *Syst Biol.*

Dornburg A, Fisk JN, Tamagnan J, Townsend JP. 2016. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. *BMC Evol Biol.* 16(1):262.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16(1): 157.

Felsenstein J. 1981. Evolutionary trees from DNA sequences—a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. *PLoS One* 7(11):e48990.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14(7):685–695.

Grama A, Karypis G, Kumar V, Gupta A. 2003. Introduction to parallel computing. Harlow (UK): Pearson Education Limited.

Gu X, Fu YX, Li WH. 1995. Maximum-likelihood-estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol.* 12(4):546–557.

Guennebaud G, Jacob B, et al. 2010. Eigen v3. Version 3. Available from: http://eigen.tuxfamily.org. Accessed February 6, 2020.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59(3):307–321.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34(23):4121–4123.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Le SV. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35(2):518–522.

Izquierdo-Carrasco F, Gagneur J, Stamatakis A. 2012. Trading memory for running time in phylogenetic likelihood computations. Bioinformatics Conference. Vilamoura, Portugal.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14(6):587–589.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21):4453–4455.

Lanfear R, Calcott B, Ho SY, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol.* 29(6):1695–1701.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol.* 29(10):2921–2936.

Le SQ, Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol.* 59(3):277–287.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc B* 363:3965–3976.

Lemey P, Salemi M, Vandamme A-M. 2009. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. New York: Cambridge University Press.
Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol*. 50(6):913–925.
Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*. 21(9):1781–1791.
Minh BQ, Hahn MW, Lanfear R. 2018. New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv* 487801, doi: 10.1101/487801.
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. 30(5):1188–1195.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17):i541–548.
Moler C, Loan CV. 1978. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev*. 20(4):801–836.
Morel B, Kozlov AM, Stamatakis A. 2019. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* 35(10):1771–1773.
Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
Norris JR. 1997. Markov chains. Cambridge: Cambridge University Press.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
Sanderson MJ, McMahon MM, Steel M. 2011. Terraces in phylogenetic tree space. *Science* 333(6041):448–450.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3):502–504.
Schrempf D, Minh BQ, De Maio N, von Haeseler A, Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J Theor Biol*. 407:362–370.
Schrempf D, Minh BQ, von Haeseler A, Kosiol C. 2019. Polymorphism-aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. *Mol Biol Evol*. 36(6):1294–1301.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51(3):492–508.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 16(8):1114–1116.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12): 1246–1247.
Snir M, Otto SW, Huss-Lederman S, Walker DW, Dongarra J. 1998. MPI: the complete reference—the MPI core. Cambridge (MA): The MIT Press.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313.
Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc Lond B* 269(1487):137–142.
Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A*. 94(13):6815–6819.
Wang HC, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol*. 67(2):216–235.
Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nat Ecol Evol*. 1(11):1737–1746.
Woodhams MD, Fernandez-Sanchez J, Sumner JG. 2015. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. *Syst Biol*. 64(4):638–650.
Yang Z. 1994a. Estimating the pattern of nucleotide substitution. *J Mol Evol*. 39(1):105–111.
Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39(3):306–314.
Zhou XF, Shen XX, Hittinger CT, Rokas A. 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol Biol Evol*. 35(2):486–503.

OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough

Hervé Philippe[1]*, Henner Brinkmann[1], Dennis V. Lavrov[2], D. Timothy J. Littlewood[3], Michael Manuel[4], Gert Wörheide[5,6], Denis Baurain[7]

1 Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada, **2** Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, United States of America, **3** Department of Zoology, The Natural History Museum, London, United Kingdom, **4** Université Paris 6, UMR 7138 "Systématique, Adaptation, Evolution" UPMC CNRS IRD MHNH, Paris, France, **5** Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, München, Germany, **6** GeoBio-Center, Ludwig-Maximilians-Universität München, München, Germany, **7** Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

In the quest to reconstruct the Tree of Life, researchers have increasingly turned to phylogenomics, the inference of phylogenetic relationships using genome-scale data (Box 1). Mesmerized by the sustained increase in sequencing throughput, many phylogeneticists entertained the hope that the incongruence frequently observed in studies using single or a few genes [1] would come to an end with the generation of large multigene datasets. Yet, as so often happens, reality has turned out to be far more complex, as three recent large-scale analyses, one published in *PLoS Biology* [2–4], make clear. The studies, which deal with the early diversification of animals, produced highly incongruent (Box 2) findings despite the use of considerable sequence data (see Figure 1). Clearly, merely adding more sequences is not enough to resolve the inconsistencies.

Here, taking these three studies as a case in point, we discuss pitfalls that the simple addition of sequences cannot avoid, and show how the observed incongruence can be largely overcome and how improved bioinformatics methods can help reveal the full potential of phylogenomics.

Hurdles to Phylogenomics

Two factors contribute significantly to the difficulty of reconstructing the correct phylogenetic tree for a set of sequences. First, if speciation events are closely spaced in time, the amount of phylogenetic signal is often small, leading to short internal tree branches that are difficult to resolve [5,6]. Second, if the events of interest are ancient, terminal branches tend to be long and replete with multiple substitutions occurring at the same position (i.e., homoplasy). In the extreme case, insufficient signal may remain for very deep divergences to be resolved even when using very long gene sequences [7]—but this issue is outside the scope of the present contribution. Depending on the accuracy of the model of sequence evolution, multiple substitutions can go undetected or be wrongly inferred. In both situations spurious phylogenetic signals are generated; these constitute the major part of what we collectively term non-phylogenetic signal. The best known example of the misleading effect of non-phylogenetic signal is the long branch attraction (LBA) artifact [8]: when two (or more) lineages have much longer branches than the others, they tend to group together irrespective of their true relationships. Notably, the outgroup is a natural source of long branches that may attract fast-evolving (hence long branched) species of the ingroup. When this happens, attracted branches artifactually emerge too deeply in the tree [9].

Inferring phylogenies in difficult cases is akin to finding a needle (phylogenetic signal) in a haystack. Under the oversimplified assumption of an absence of non-phylogenetic signal, one can compute that the resolving power would increase from approximately 15 million years when using small subunit ribosomal RNA alone to less than 1 million years when using more than 50 genes [10]. At such levels of resolution, incomplete lineage sorting (i.e., the retention of ancestral polymorphisms over successive speciation events) should be taken into account as a potential source of phylogenetic error [11]. Nonetheless, even if conflicting gene genealogies were not an issue, throwing additional gene sequences at a difficult phylogenetic question does not necessarily solve the problem—the size of the needle is indeed increased, but so too is the size of the haystack. It follows that non-phylogenetic signal may become dominant and yield incongruent, yet statistically highly supported, phylogenomic trees [12].

How to Prevent Deleterious Effects of Non-Phylogenetic Signal

Non-phylogenetic signal has multiple and disparate sources [13]. When multiple genes are concatenated and analyzed with standard methods (but see [14]), non-phylogenetic signal is caused by the inclusion of sequences that deviate from the true species phylogeny or by the inability of our methods to correctly handle multiple substitutions. In practice, it mainly stems from (i) the incorrect identification of orthologs, (ii) erroneous alignments, or (iii) the incorrect reconstruction of multiple substitutions occurring at a given position, the last owing to model violations in probabilistic methods

The Perspective section provides experts with a forum to comment on topical or controversial issues of broad interest.

Citation: Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol 9(3): e1000602. doi:10.1371/journal.pbio.1000602

Academic Editor: David Penny, Massey University, New Zealand

Published March 15, 2011

Copyright: © 2011 Philippe et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The work was funded by NSERC (www.nserc-crsng.gc.ca), CRC (www.chairs-chaires.gc.ca), Agence Nationale de la Recherche (http://www.agence-nationale-recherche.fr/), ARC Biomod (www.cfwb.be), and DFG (http://www.dfg.de/en/index.jsp). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: BS, bootstrap support; EST, expressed sequence tag; LBA, long branch attraction

* E-mail: herve.philippe@umontreal.ca

Box 1. From Phylogenetics to Phylogenomics

Phylogenetics, the determination of evolutionary relationships among organisms, is central to our understanding of the evolution of life. For instance, the three phylogenies of Figure 1 entail profoundly different interpretations about the complexity of the common ancestor of all animals. Important body plan characters (e.g., neurosensory and digestive systems and muscle cells) are found in cnidarians, ctenophores, and bilaterians but not in sponges and placozoans. According to the phylogenies of Schierwater et al. [4] and Dunn et al. [2], the taxonomic distribution of these characters implies either (i) that the ancestral metazoan already featured these traits and that sponges (and placozoans) have secondarily lost them or (ii) that these characters were acquired several times independently by convergence (e.g., in the cnidarian + ctenophore and in the bilaterian lineages, according to the tree of Figure 1A). In contrast, the phylogeny of Philippe et al. [3] is more congruent with morphological characters and compatible with a simple metazoan ancestor and a later emergence of these characters only once, in the lineage leading to the common ancestor of coelenterates (cnidarians+ctenophores) and bilaterians.

Phylogenies are generally depicted as trees (which are non-reticulated graphs, as in Figure 1) because vertical evolution is undisputedly the primary mechanism of inheritance for genetic material. However, the existence of horizontal transmission (e.g., hybridization of closely related taxa, organelle acquisition through endosymbiosis and horizontal gene transfer) makes phylogenetic trees only pragmatic approximations, which will probably be replaced by phylogenetic networks in the long term (particularly for unicellular organisms).

Recently, phylogenomics, the use of genomic data to infer evolutionary relationships, has emerged as a new domain of phylogenetics. The main strength of phylogenomics is the drastic reduction in random (or sampling) error brought by the use of large (multigene) datasets. Numerous approaches can be used to take advantage of genomic data (for review see [49]). Briefly, new methods based on oligonucleotide content, gene content, or intron positions look promising (as shown by their ability to yield reasonable trees) but require additional theoretical developments to achieve their full potential. That is why the two most popular phylogenomic approaches are simple extensions of the standard phylogenetics methods applied to single-gene datasets. The first, known as the "supermatrix" (or superalignment), consists in concatenating numerous orthologous genes into a single supergene, which is analyzed using standard methods (or slightly modified methods such as separate models allowing for multiple sets of branch lengths [50]). The second, "supertree," approach takes the opposite path by first inferring a tree for each gene in the dataset and then combining these individual trees into a single supertree. The supermatrix approach is the most commonly used, in agreement with the handful of studies suggesting that it offers greater accuracy than the supertree [13,51], though this remains to be formally demonstrated.

(i.e., Bayesian inference and maximum likelihood). Although all three aspects have received considerable attention from theoreticians, and despite the availability of numerous bioinformatics tools [15–17], there is still no magic bullet. That is why classic phylogenetics involves numerous refinements and controls, which are difficult, but not impossible, to apply at a phylogenomic scale.

Non-phylogenetic signal can be reduced by improving (i) the quality of primary alignments through selection of the orthologous genes that are least subject to saturation and (ii) the detection of multiple substitutions, which is best achieved by using both a large number of species and the most realistic model of sequence evolution. In the following, we show that both improvements are required at the same time to address the difficult question of the relationships among major animal groups, i.e., sponges, placozoans, ctenophores, cnidarians, and bilaterians. Reanalysis of the underlying data indicates that failure to apply one or more of the strategies intended to decrease non-phylogenetic signal is what caused the incongruent, though strongly supported, results that were recently observed [2–4].

Issues at the Level of Sequence Alignments

Selection of unambiguously orthologous genes [18] is usually achieved by targeting single-copy genes (e.g., mitochondrial genes) or pre-selected genes (e.g., ribosomal RNAs and proteins), or through automatic clustering methods. None of these options are without problems. Both manual and automatic methods [19–22] heavily rely on BLAST similarity scores, which are known to be a poor estimator of the true evolutionary distance [23]. Given the limitations of existing methods of orthology detection (Box 3), careful phylogenetic analysis of each alignment is important to achieve maximal accuracy. However, this manual step is difficult and subjective. That is why it is preferable to also verify orthology a posteriori. One possibility is to assess whether branches receiving high statistical support from every single gene tree are congruent with the species tree [18]. Though the latter is unknown, the phylogeny obtained by the concatenation of numerous genes constitutes a reasonable approximation. Hence, Philippe et al. [3] looked at every supported branch (bootstrap support [BS]$\geq$70%) from single-gene trees that were incongruent with the concatenated tree to assess the orthology of their preselected genes. Only 6.5% of the branches were incongruent, and almost all conflicts were best explained by reconstruction errors affecting single-gene trees [3]. According to this semi-automated approach, the 128 genes used in [3] can be provisionally considered as orthologous and suitable for phylogenetic analysis. In contrast, when applied to the datasets of Schierwater et al. [4] and Dunn et al. [2], the very same approach identifies several instances of incongruence between single-gene and concatenated trees (mainly apparent horizontal gene transfers that are in fact more likely due to contaminations, or deep unrecognized paralogy; see Text S1 and Figures S1, S2, S3, S4, S5, S6, S7, S8, S9).

This as well as the discovery of other important issues (see Table S1) prompted us to reassess and reanalyze the dataset of Schierwater et al. [4]. The revised phylogeny we generate (Figure 2B) differs from the original one (Figure 2A) in the deep animal relationships: the strong support for a sister-group relationship between Bilateria and a group composed of placozoans, sponges (Porifera), ctenophores, and cnidarians [4] has vanished, and sponges are now recovered as the sister group of all other Metazoa. Strikingly, this part of the revised tree (Figure 2B) suffers from a lack of statistical support (all BS$<$50% except for the monophyly of cnidarians). The simplest explanation for these results (Figure 2B) is that the genuine phylogenetic signal for non-bilaterian animal relationships is scarce, as reported in all previous studies (e.g., [24–28]). The possible inclusion of non-orthologous sequences (see Figures S1,

Box 2. Glossary

Homology/orthology/paralogy/xenology: Genes that derive from a common ancestor are termed homologs. Two homologous genes are orthologous if they diverged through a speciation event. In contrast, paralogs originate by duplication of a single gene within a given lineage, whereas xenologs result from the horizontal transfer of a gene from a donor species to a receiver species (which might eventually get its original copy replaced by the xenolog).

Homoplasy/convergence: Spurious similarity due to convergence or reversion and not to common ancestry is termed homoplasy. Convergence describes the independent acquisition by separate evolutionary lineages of the same nucleotide (or amino acid) at a given position. This is a direct consequence of multiple substitutions.

Incomplete lineage sorting: The transient retention of ancestral polymorphisms across speciation events. Speciations compressed in time and large reproductive populations both increase the likelihood of this phenomenon. Considering three lineages having rapidly diverged, by chance some sequence positions will be shared between one pair, while others will be shared between another pair, and yet others between the third possible pair, hence blurring the phylogenetic signal on the corresponding branches.

Incongruence: Two (or more) phylogenetic trees are said to be incongruent when they exhibit conflicting branching orders (i.e., topologies) and cannot be superimposed. This implies that at least one node (also known as a bipartition) present in one tree is not found in the other(s), where it is replaced by alternative groupings of taxa.

Model of sequence evolution: A statistical description of the process of substitution in nucleotide or amino acid sequences. Complex models better approximate the evolutionary process but at the expense of more parameters and computational time. As parameter-rich models require more data to behave properly, they have become really useful with the advent of phylogenomic datasets.

Monophyly: To be considered monophyletic, a taxonomic group must satisfy two conditions: (i) all its taxa must derive from a single ancestor and, reciprocally, (ii) all taxa deriving from this common ancestor must belong to the group.

Non-phylogenetic signal: The combination of different kinds of structured noise (e.g., undetected homoplasies) that compete with the genuine phylogenetic signal during tree reconstruction. Even if the non-phylogenetic content is partly a property of a multiple sequence alignment (notably related to its saturation level), the non-phylogenetic signal actually inferred heavily depends on the method and the model of evolution selected. In probabilistic methods, the non-phylogenetic signal mainly results from the data violating the model of sequence evolution. These violations arise because our models are inevitably oversimplified in comparison to the complexity of the natural evolutionary process. Eventually, the apparent signal analyzed will be a blend of phylogenetic and non-phylogenetic signal.

Outgroup/ingroup: Nearly all tree reconstruction methods produce unrooted trees, in which inferred relationships do not convey any information about the direction of time. To root a tree and turn it into a phylogeny, one has to include in the analysis a group of taxa that are known to be outside the group under study. This reference group is termed the outgroup, while the taxa of interest make the ingroup.

Patristic distance: The sum of the lengths of the branches that connect two nodes in a phylogenetic tree, where those nodes are typically terminal nodes representing extant taxa. It is thus an inferred distance (taking into account multiple substitutions) greater than the uncorrected distance directly computed from the number of differences observed between the two corresponding sequences in the alignment.

Phylogenetic signal/synapomorphy: The substitutions occurring along a given branch of the evolutionary tree. The strength of the phylogenetic signal is proportional to the number of substitutions occurring along the branch. In non-probabilistic methods, the signal is encoded in synapomorphies, i.e., shared residues (nucleotides or amino acids) at aligned positions that are specific to a set of sequences derived from a common ancestor. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to increase with the fit of the model to the data (i.e., the ability of the model to explain the data).

Phylogenetic tree: A (connected acyclic) graph describing the estimated evolutionary relationships among a group of species. In molecular trees, branch lengths are proportional to the genetic distances (and hence to some extent to time) inferred from the analysis of a multiple alignment of homologous sequences (nucleotide or amino acid sequences).

Probabilistic methods: A family of tree reconstruction methods from multiple sequence alignments that are grounded in statistical theory and make use of explicit models of sequence evolution. These include maximum likelihood and Bayesian inference approaches and are known to be the most accurate but also the most computationally demanding.

Saturation: When sequences in a multiple alignment have undergone so many multiple substitutions that apparent distances largely underestimate the real genetic distances, the alignment is said to be saturated. Phylogenetic inference works best with datasets that are only slightly saturated. Owing to their reduced state space (four possible bases), nucleotide sequences saturate more rapidly than protein sequences (20 possible amino acids).

Site-homogeneous/site-heterogeneous models: Most models of sequence evolution assume that the same evolutionary process takes place at every position (or site) of an alignment. With such models, only the evolutionary rate can be modeled as heterogeneous across sites, usually through a gamma distribution of rates. However, selective constraints are known to be quite heterogeneous across positions, hence seriously violating the hypotheses of site-homogeneous models. On the other hand, site-heterogeneous models assume that the evolutionary process varies widely across sites, in particular the set of acceptable amino acids (e.g., in the CAT model). A number of studies have demonstrated that site-heterogeneous models provide a better fit to phylogenomic datasets and tend to reduce the sensitivity to tree reconstruction artifacts (e.g., LBA).

S2, S3, S4, S5, S6, S7, S8, S9) might create a strong signal that could overcome the genuine but faint phylogenetic signal, and lead to the incorrect—but strongly supported—monophyly of "diploblasts" (sponges+placozoans+ctenophores+cnidarians) that was observed in the original study (Figure 2A). Otherwise, the topology we infer from the revised alignments is similar to the published tree [4], with only three nodes differing out of 21. This demonstrates that phylogenomics is relatively robust to the possible inclusion of non-orthologous sequences when the genuine phylogenetic signal is abundant (see also [29,30]), which can be explained by the randomness of most of the introduced errors preventing the appearance of a structured misleading signal.

On the other hand, phylogenomics is sensitive to the non-phylogenetic signal that stems from the incorrect inference of multiple substitutions. By devoting a large part of their dataset to mitochondrial genomes, which are fast-evolving in Bilateria (e.g., [24,31]), Schierwater et al.'s solution unwittingly favored the emergence of Bilateria between the outgroup and a group composed of all the non-bilaterian Metazoa, because of the LBA artifact. This artifact probably also affects the phylogeny of Dunn et al. [2]; in that case, the fast-evolving ctenophores are likely attracted by the distant outgroup (see Text S1). In the phylogeny inferred from an updated version of the alignments of Dunn et al. (purged of several sequencing errors and species misidentifications—see Table S2—and completed with new sequences, thereby reducing the amount of missing data from 55% to 35%), sponges are the sister group of all other Metazoa, with the fast-evolving Ctenophora representing the sister group of Cnidaria plus Bilateria (Figure S11; see also [32]).

In summary, analyzing the revised alignments from Schierwater et al. [4] and Dunn et al. [2] with their original taxon sampling and inference methods is sufficient to eliminate all significant incongruences among the three recent phylogenomic studies (Figure 1). The variability in robustness across the tree (e.g., Figure 2) underscores the importance of clean phylogenomic datasets: whereas large amounts of phylogenetic signal usually drown out any non-phylogenetic signal, for nodes characterized by a scarce phylogenetic signal, even small amounts of non-phylogenetic signal may dominate and eventually yield incorrect results [10].

Issues at the Level of Taxon Sampling

The lack of support observed in Figures 2 and S11 contrasts with the high bootstrap values obtained by Philippe

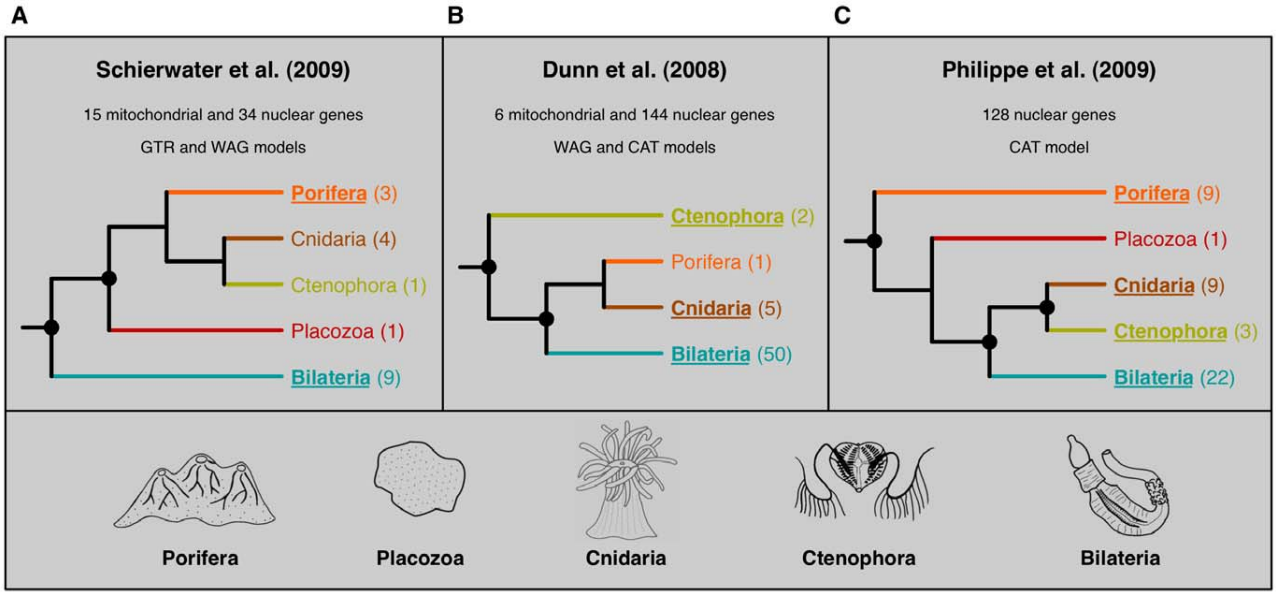


Figure 1. Simplified representation of the trees obtained in three recent phylogenomic analyses of early animal diversification. (A) Schierwater et al. [4] tree. (B) Dunn et al. [2] tree. (C) Philippe et al. [3] tree. Numbers in parentheses after taxon names indicate the number of species included in the dataset for the corresponding taxon. Bootstrap support values above 90% are indicated by a bullet (for nodes) or by underlining (for terminal taxa). It is worth mentioning that the monophyly of Porifera is not unequivocally accepted [28,46]; only the analysis of 30,000 positions with a rich taxon sampling and a complex model of evolution recovers it with significant statistical support [3]. Although such a sparse phylogenetic signal will require harnessing the full potential of phylogenomics to be confidently solved, this question is outside the scope of this study. Simplified drawings (redrawn from [74]) on the bottom illustrate the huge morphological disparity existing between the five terminal taxa. Porifera correspond to sponges; Cnidaria to sea anemones, jellyfishes, and allies; Ctenophora to comb jellies; and Bilateria to all other animals (characterized by their bilateral symmetry) except *Trichoplax* (Placozoa), which appears to be morphologically the most simply organized animal phylum.
doi:10.1371/journal.pbio.1000602.g001

Box 3. Quality Control of Phylogenomic Datasets

Despite the great progress in software development [19–22,52–54], our nine years of experience with large-scale multigene analyses [55] leads us to conclude that computer-assisted manual expertise is not yet dispensable. In particular when processing EST data, two issues are still challenging to handle by automation: (i) the non-homology of short sequence stretches due to frameshifts and point mutations and (ii) the non-orthology of one or more genes with similar sequence for some species, because of paralogy or xenology, along with taxonomic misidentifications and library contaminations (e.g., by parasites such as platyhelminthes). An important limitation of automated methods for checking single-gene alignments for orthology prior to concatenation is the limited amount of sequence information available in a single gene, which often makes current statistical analyses impractical. If the threshold used is stringent, almost every sequence will fail the test, whereas a loose threshold will lead to numerous false positives. Manual verification, through visual inspection of alignments and phylogenies, can to a large extent compensate for this lack of statistical power if a large number of species (much more than those eventually included in the final analysis) is taken into account. First, as conserved positions are clearly identified, both translational frameshifts (leading to stretches of amino acids highly different from the consensus, which are mostly found at EST extremities) and local sequencing errors (visible as unmatched amino acids at highly conserved positions) stand out. Based on manual analysis, we estimate that approximately 4,800 amino acids (0.66% of the complete alignment) were erroneous in the Dunn et al. dataset [2] because of frameshifts and local sequencing errors (including incorrect translation owing to a mistake in the specification of the genetic code for ambulacrarian mitochondria; see Table S2). Second, xenology, contaminations, and misidentification can be efficiently detected when individual alignments encompass a broad taxonomic diversity, as such diversity is much more likely to find a close relative of the donor species. For instance, in the Dunn et al. dataset [2], one acoel species, the marine flatworm *Neochildia fusca*, was contaminated by microsporidia (see Table S2). Since original alignments lacked microsporidial sequences, the contamination was overlooked and acoel sequences were simply considered as extremely divergent. Similarly, hidden paralogy is easier to detect with numerous species on hand (and with deeper sequencing of each of them), because they increase the chance of finding a species that has kept both copies. Interestingly, much more serious errors (including the use of paralogous, rather than orthologous, copies, and taxonomic misidentification; see Figures S1, S2, S3, S4, S5, S6, S7, S8, S9) were identified in the manually assembled Schierwater et al. dataset [4] than in the automatically assembled Dunn et al. dataset [2] (compare Tables S1 and S2). Manual assessment of the quality of primary data is particularly tedious and time-consuming, as well as error-prone. That is why automated approaches featuring refined statistics (e.g., hidden Markov models detecting frameshifts) are strongly needed to both speed up and improve the construction of phylogenomic datasets. Finally, it should be noted that missing data (i.e., incomplete sequences), which are on the rise in recent large-scale analyses (e.g., 55.5% of the characters in [2] and 81% in [46]), constitute an additional unpredictable issue, as they might further erode statistical power and sometimes enhance tree reconstruction artifacts [38,42] (see Text S1 and Figure S11).

et al. [3] for the monophyly of each of the Porifera (96%), Coelenterata (Cnidaria+Ctenophora, 93%), and Eumetazoa (all animals except Porifera and Placozoa, 90%) (Figure 3A). However, the number of non-bilaterian metazoan species used in [3] is larger, 22 versus 9 [2,4], which could account for the difference. Indeed, it is well known that including more species allows for a better detection of multiple substitutions [33], as it decreases the amount of non-phylogenetic signal while preserving phylogenetic signal [34]; this is why authors often mention that their results should be viewed as provisional until more taxa are considered (e.g., the position of Ctenophora in [2]). To test this hypothesis, we reduced the taxon sampling of [3] to match as closely as possible the sampling of Figure 2. Even though sequences and inference methods are exactly as in [3], the support for deep animal relationships decreases drastically (Figure 3B). While the monophyly of each of the Cnidaria (94%), Coelenterata (70%), and Demospongiae + Hexactinellida (86%) still receive some support, remaining relationships are unresolved (BS<60%); in particular, Porifera and Eumetazoa are not recovered. These results corroborate the hypothesis that the use of a limited number of species generates enough non-phylogenetic signal to swamp most of the faint genuine phylogenetic signal present in this part of the animal phylogeny (owing to short internal branches and heterogeneous rates among species).

However, taxon sampling is not simply a matter of number of species [35–37]. In particular, the inclusion of both slowly evolving species and closely related outgroups (e.g., choanoflagellates for animals; see [3] and Text S1) is often of prime importance. This point is well illustrated by a reanalysis of the original alignments of Schierwater et al. in which we eliminated the most distant outgroups. When rooting exclusively with choanoflagellates, the bootstrap support for a position of Porifera as the sister group to remaining animals rises to 80% (Figure S13). Although discarding very distant outgroups (e.g., Bacteria) undoubtedly improves accuracy, the effect of including moderately distant outgroups (e.g., Fungi) in addition to close outgroups (e.g., choanoflagellates) is more difficult to assess. Eventually, it will depend on the relative influence of introducing a very long branch (the distant outgroup) and breaking up an already existing long branch (the close outgroup). Even if further studies are needed to clarify this point, an effort to increase the taxon sampling of the close outgroup should help to resolve deep animal relationships.

Finally, phylogenomic datasets, especially when based on expressed sequence tag (EST) data, are frequently characterized by incomplete gene coverage for some taxa. Yet, there have been few attempts to determine whether missing data per se can cause errors in tree reconstruction [36,38–42] and how they may interfere with other aspects of phylogenetic inference. In particular, it is not known whether a smaller, but complete, alignment of targeted genes (e.g., selectively amplified by PCR) would yield a more accurate and robust tree than a large, but incomplete, alignment of highly expressed genes (obtained by EST sequencing). These questions can and should be better assessed in the near future.

Issues at the Level of Tree Reconstruction Methods

To further explore the idea that the paramount issue in phylogenomics pertains to the reduction of non-phylogenetic signal (more than the increase of phylogenetic signal with datasets containing more and more genes, especially in the short run), we

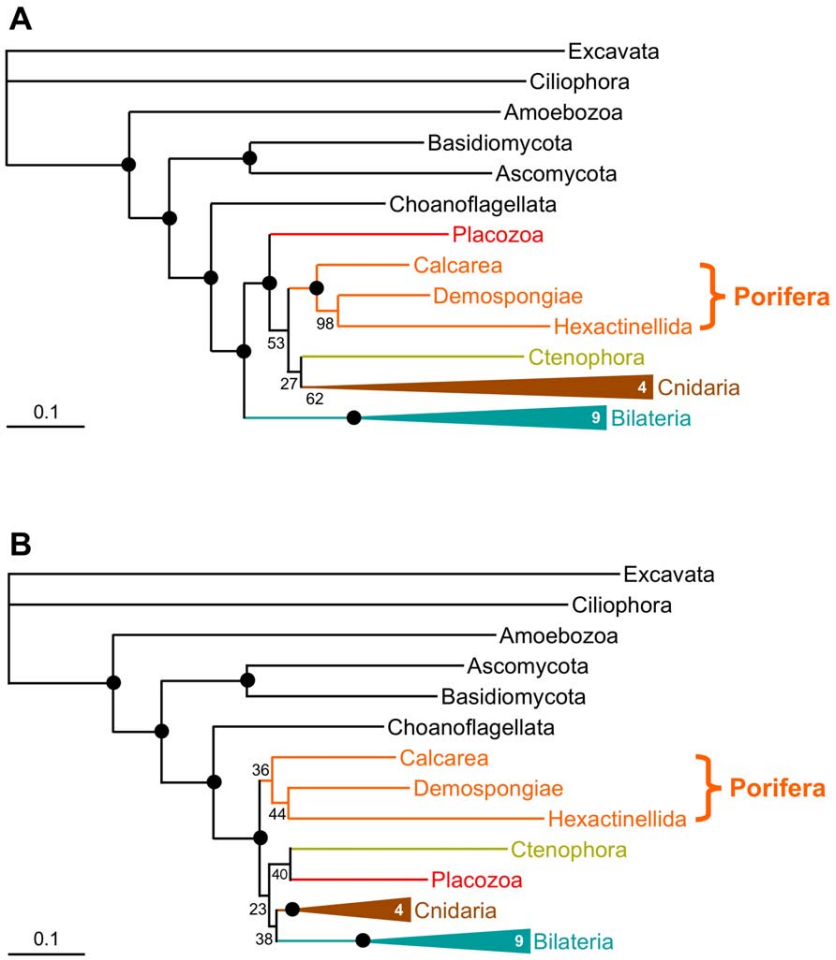


Figure 2. Analysis of the revised Schierwater et al. dataset. (A) Scheme of the original tree [4]. (B) Scheme of the tree obtained with the revised dataset. Both trees were inferred using exactly the same probabilistic method and model (i.e., using RAxML [75] with a GTR+Γ model for nucleotide sequences and a LG+F+Γ model for protein sequences). Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. In the revised dataset, numerous discrepancies were corrected (Table S1), and a few genes were discarded because of dubious orthology; 14,112 unambiguously aligned positions were retained. Furthermore, the erroneous use of mitochondrial sequences of demosponge origin to represent both hexactinellids and calcareans (Figure S9) in the original study [4] drastically—yet probably artifactually—strengthened the support for the monophyly of sponges (BS = 100%; [A]), whereas it appeared much weaker in our reanalysis (BS = 36%; [B]), in line with previous studies [24,26–28] that failed to find significant support for or against sponge monophyly (but see [3]). See Figure S10 for the complete tree obtained with the revised dataset.
doi:10.1371/journal.pbio.1000602.g002

now turn to the selection of the model of sequence evolution. Since their origin [43], the main objective of these models has been to efficiently detect multiple substitutions (Box 4). We reanalyzed the dataset of [3] with a less accurate model, i.e., the site-homogeneous WAG+F+Γ model [44] used in [4] instead of the site-heterogeneous CAT+Γ model [45] used in the original study [3] (Figure 4A). In the WAG+F+Γ tree (Figure 4B), not only does resolution decrease (see BS of 43%, 45%, or 55%), but also the fast-evolving ctenophores now emerge at the base of all animals with strong support (BS = 98%), exactly as expected for a LBA artifact due to model mis-specifications. This indicates that when the less appropriate WAG+F+Γ model is used, multiple substitutions are so poorly inferred that branch lengths are miscalculated (i.e., non-phylogenetic signal has overwhelmed phylogenetic signal).

In summary, the incongruence at the base of the animal tree observed in recent phylogenomic studies [2–4] can be explained by (i) a limited amount of phylogenetic signal, reflected in the short internal branches, and (ii) a profusion of confounding non-phylogenetic signal in certain cases. Since genuine phylogenetic signal is similar in all three analyses (i.e., internal branch lengths are identical and datasets are of similar size), conflicts are due to variations in the level of non-phylogenetic signal—depending on the quantity of non-orthologous sequences included, the number of species considered, and the model of sequence evolution selected. Ultimately, the ratio of phylogenetic to non-phylogenetic signal will determine the outcome: (i) when the phylogenetic signal is strong (sufficiently long internal branches), phylogenomics is always able to recover the correct topology, as found in the three studies [2–4] for outgroup and bilaterian phylogenies; (ii) when both signals are weak, results are statistically non-significant, as is often observed for deep animal relationships; and (iii) when the phylogenetic signal is weak (short internal branches) and the non-phylogenetic signal is strong (e.g., scarce taxon sampling), an artifactual topology is robustly inferred, such as the monophyly of "diploblasts" [4] or the basal

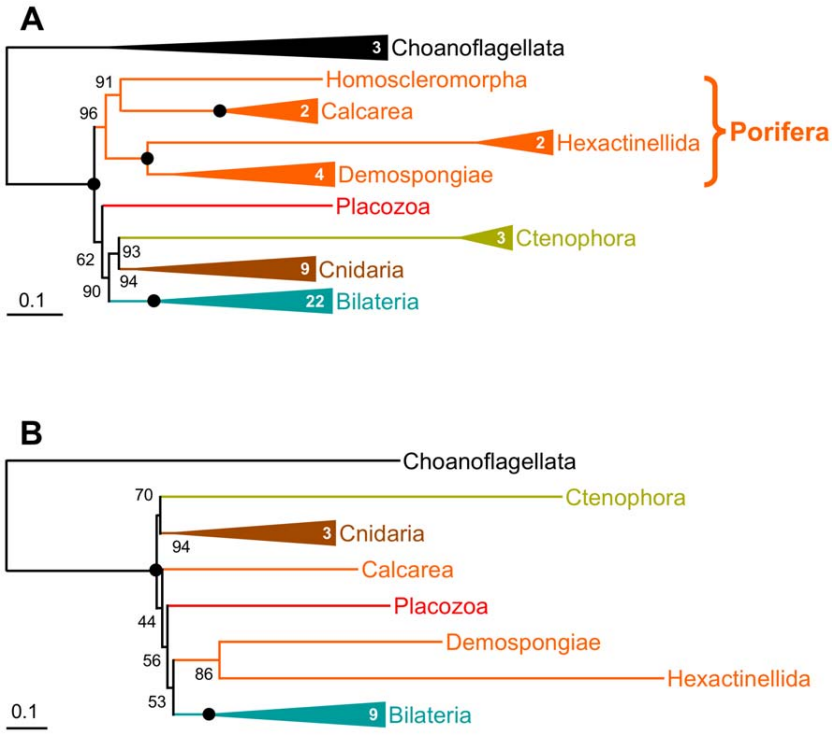


Figure 3. Reanalysis of the Philippe et al. dataset with a reduced taxon sampling. (A) Scheme of the original tree [3]. (B) Scheme of the tree obtained after reduction of the taxon sampling. Both trees were inferred using exactly the same probabilistic method and model (i.e., PhyloBayes using the CAT+Γ model [76]). Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. See Figure S12 for the complete tree obtained after reduction of the taxon sampling.
doi:10.1371/journal.pbio.1000602.g003

Box 4. Improving Phylogenetic Inference Methods

There is broad consensus on the necessity of using probabilistic methods in phylogenetic inference. Development of more accurate models of sequence evolution is central to the improvement of these methods. This generally implies more complex models, which are expected to come with increased computational load. Hence, in-depth analyses of datasets that are rich in both genes and species with such models can become prohibitive [46]. Consequently, some promising approaches, e.g., accounting for three-dimensional structure of proteins [56,57] or performing joint alignment and phylogeny [58,59], will probably stay out of reach for years. Fortunately, numerous recent algorithmic developments [60–62] significantly speed up phylogenetic computations, thus paving the way for model improvements. One generally considers that models should be biologically sound. Although biological realism is particularly important for understanding molecular evolution, it is less central for phylogenetic inference, where improving detection of multiple substitutions should be the top priority. As a result, models that more accurately distinguish a synapomorphy from a convergence greatly improve phylogenetic accuracy. Briefly, major steps forward were the modeling of heterogeneity of rate across species [63], heterogeneity of rate across substitutions [64,65], heterogeneity of nucleotide/amino acid composition across species [66,67], heterogeneity of rate across sites [68], and heterogeneity of the substitution process across sites [45]. In contrast, some other improvements, e.g., to handle heterotachy (i.e., heterogeneity of rate over time), had limited effects on phylogenetic reconstruction [69]; heterogeneity of rates across genes, handled by separate models [50], also has limited impact ([70], but see [71]). Future progress is expected (i) from the combination of various existing models [72], (ii) from the handling of other complexities, such as the heterogeneity of the substitution process over time, and (iii) from the handling of incomplete lineage sorting [11,73].

emergence of ctenophores (Figure 4B) (see also [2,32,46]).

Issues at the Level of Gene Sampling

Last but not least, it should be noted that not all genes contain the same potential amount of non-phylogenetic signal. Depending on both functional constraints and evolutionary trajectory, different genes can include positions subject to different ranges of multiple substitutions, i.e., they may display variable levels of saturation. To estimate the saturation in the three datasets [2–4], we used the comparison of patristic and uncorrected distances [47]. As shown by the slope of the regression line (data without any saturation have slope = 1; see [12]), the three datasets (Figure 5) are different, with that of Schierwater et al. being the most saturated (slope = 0.38) and that of Philippe et al. the least affected by multiple substitutions (slope = 0.53). This uneven amount of non-phylogenetic signal explains in part the differences observed in the three studies, but is difficult to separate from other factors. The phylogeny of Figure 1C, with the monophyly of each of Coelenterata (cnidarians+ctenophores) and

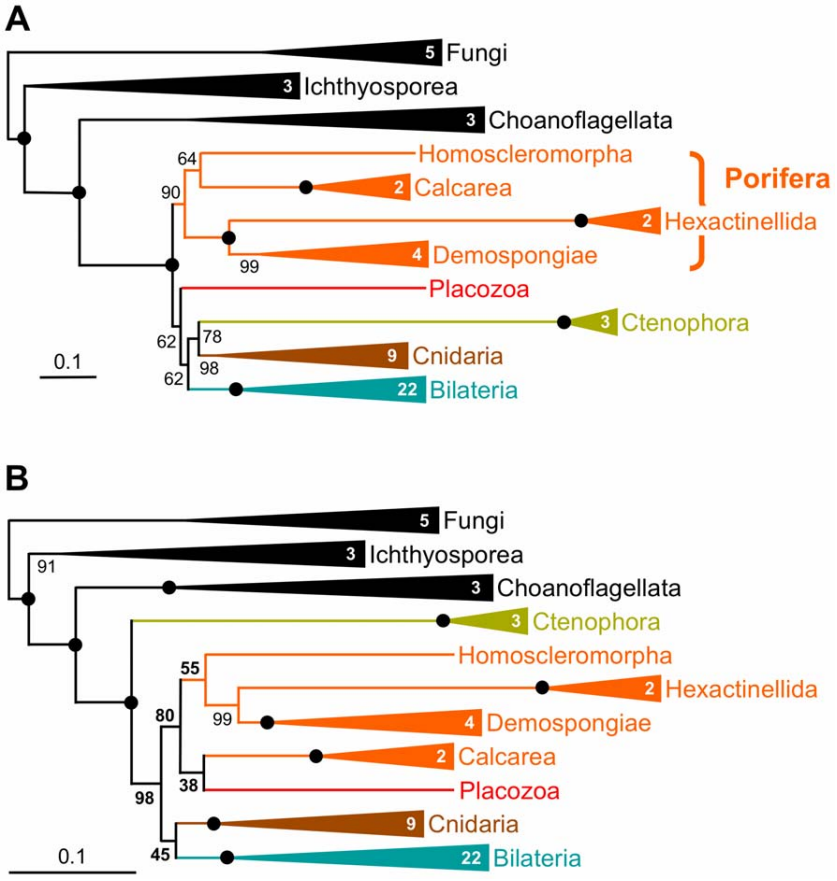


Figure 4. Reanalysis of the Philippe et al. dataset with a less complex model. (A) Scheme of the original tree [3] obtained with the CAT+Γ model. (B) Scheme of the tree obtained with the less complex WAG+F+Γ model. Both trees were inferred using exactly the same dataset. The WAG+F+Γ model has a less good fit to this alignment than the CAT+Γ model [3]. Numbers in the triangles indicate the number of species used for the corresponding clade. Bullets denote maximum bootstrap support values (BS = 100%); lower values are given. See Figure S14 for the complete tree obtained with the less complex WAG+F+Γ model.
doi:10.1371/journal.pbio.1000602.g004

Eumetazoa (all animals except sponges and placozoans), could be considered as the working hypothesis, because Philippe et al. [3] strived to minimize all three sources of non-phylogenetic signal (through the use of weakly saturated genes, a large number of species, and a complex model of sequence evolution). Nevertheless, the scarcity of phylogenetic signal shown here argues strongly for additional studies to confidently resolve the relationships among non-bilaterian animals.

Conclusion

Contrary to common belief, some degree of conflict has to be expected when applying phylogenomics to difficult phylogenetic questions, because of the preva-

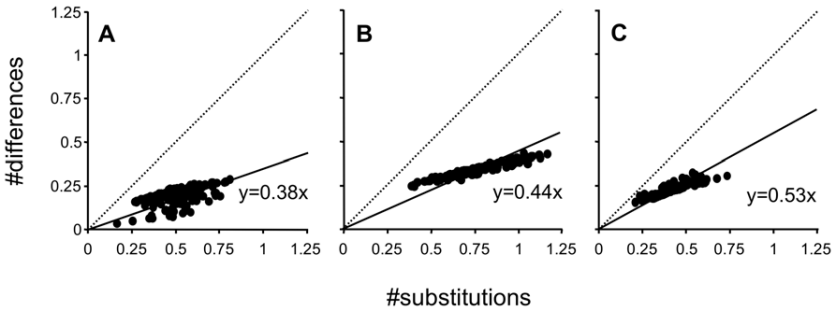


Figure 5. Saturation levels of datasets from Schierwater et al., Dunn et al., and Philippe et al. (A) Schierwater et al. [4] dataset. (B) Dunn et al. [2] dataset. (C) Philippe et al. [3] dataset. The revised alignments from Schierwater et al. and Dunn et al. were used (available as Datasets S1 and S2; see Text S1). The level of saturation was estimated for each dataset by computing the slope of the regression line of patristic distances (*y*-axis) versus uncorrected distances (*x*-axis), as previously described [12]. Patristic distances between two species were computed from branch lengths of the best maximum likelihood tree (using a GTR+Γ model for nucleotide sequences and a LG+F+Γ model for protein sequences).
doi:10.1371/journal.pbio.1000602.g005

lence of non-phylogenetic signal. Consequently, we stress the necessity of reducing its impact. Since taxon and gene sampling is being rapidly improved by the relentless progress in sequencing technology (even if obtaining well preserved and correctly identified specimens remains the limiting factor for several key taxa), full achievement of the ultimate goal of phylogenomics—i.e., accurate resolution of the Tree of Life—will primarily hinge on better procedures for the selection of orthologous and least saturated genes as well as on improved models of sequence evolution. In summary, while we certainly encourage the inclusion of neglected groups of organisms in large-scale sequencing studies (e.g., [2,3,46,48]), we consider at least as important that phylogeneticists engage in theoretical and bioinformatics developments that keep pace with sequencing technology to overcome these serious bottlenecks. This is essential to ensure that lessons learned from classical and molecular systematics are not forgotten in the phylogenomic era.

Supporting Information

Dataset S1 Updated alignment of the Schierwater et al. dataset under the Nexus format.
Found at: doi:10.1371/journal.pbio.1000602.s001 (0.10 MB ZIP)

Dataset S2 Updated alignment of the Dunn et al. dataset under the Nexus format.
Found at: doi:10.1371/journal.pbio.1000602.s002 (0.46 MB ZIP)

Figure S1 Phylogeny of the AT6 gene.
Found at: doi:10.1371/journal.pbio.1000602.s003 (0.06 MB PDF)

Figure S2 Phylogeny of the CDC gene.
Found at: doi:10.1371/journal.pbio.1000602.s004 (0.06 MB PDF)

Figure S3 Phylogeny of the RP3 gene.
Found at: doi:10.1371/journal.pbio.1000602.s005 (0.06 MB PDF)

Figure S4 Phylogeny of the EF1 gene.
Found at: doi:10.1371/journal.pbio.1000602.s006 (0.04 MB PDF)

Figure S5 Phylogeny of the H70 gene.
Found at: doi:10.1371/journal.pbio.1000602.s007 (0.04 MB PDF)

Figure S6 Phylogeny of the PAX gene.
Found at: doi:10.1371/journal.pbio.1000602.s008 (0.04 MB PDF)

Figure S7 Phylogeny of the RAS gene.
Found at: doi:10.1371/journal.pbio.1000602.s009 (0.06 MB PDF)

Figure S8 Phylogeny of the CO2 gene.
Found at: doi:10.1371/journal.pbio.1000602.s010 (0.06 MB PDF)

Figure S9 Taxonomic misidentification for mitochondrial proteins of sponges.
Found at: doi:10.1371/journal.pbio.1000602.s011 (0.06 MB PDF)

Figure S10 Analysis of the revised Schierwater et al. dataset.
Found at: doi:10.1371/journal.pbio.1000602.s012 (0.04 MB PDF)

Figure S11 Analysis of the updated Dunn et al. dataset.
Found at: doi:10.1371/journal.pbio.1000602.s013 (0.07 MB PDF)

Figure S12 Reanalysis of the Philippe et al. dataset with a reduced taxon sampling.
Found at: doi:10.1371/journal.pbio.1000602.s014 (0.04 MB PDF)

Figure S13 Reanalysis of the original Schierwater et al. alignment with only the closest outgroup (Choanoflagellata).
Found at: doi:10.1371/journal.pbio.1000602.s015 (0.04 MB PDF)

Figure S14 Reanalysis of the Philippe et al. dataset with a less complex model.
Found at: doi:10.1371/journal.pbio.1000602.s016 (0.06 MB PDF)

Table S1 List of errors detected in the dataset of Schierwater et al.
Found at: doi:10.1371/journal.pbio.1000602.s017 (0.05 MB PDF)

Table S2 List of errors detected in the dataset of Dunn et al.
Found at: doi:10.1371/journal.pbio.1000602.s018 (0.13 MB PDF)

Text S1 Methods and supporting information.
Found at: doi:10.1371/journal.pbio.1000602.s019 (0.16 MB DOC)

Acknowledgments

We thank Casey Dunn, Gavin Naylor, Davide Pisani, Scott Roy, Bernd Schierwater, and Mike Steel for critical comments on earlier versions of the manuscript. G. W. is supported by the Deutsche Forschungsgemeinschaft through the Priority Program SPP1174 "Deep Metazoan Phylogeny," D. B. by the Communauté Française de Belgique (ARC Biomod), M. M. by the Agence Nationale de la Recherche, and H. P. by the Canadian Research Chair program and Natural Sciences and Engineering Research Council. The Réseau Québécois de Calcul Haute Performance provided computational resources.

References

1. Gee H (2003) Evolution: ending incongruence. Nature 425: 782.
2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.
3. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19: 706–712.
4. Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, et al. (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol 7: e1000020. doi:10.1371/journal.pbio.1000020.
5. Philippe H, Chenuil A, Adoutte A (1994) Can the Cambrian explosion be inferred through molecular phylogeny? Development 120: S15–S25.
6. Saitou N, Nei M (1986) The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol 24: 189–204.
7. Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: Gascuel O, ed. Mathematics of evolution and phylogeny. Oxford: Oxford University Press. pp 384–412.
8. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410.
9. Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8: 616–623.
10. Baurain D, Philippe H (2010) Current approaches to phylogenomic reconstruction. In: Caetano-Anollés G, ed. Evolutionary genomics and systems biology. Hoboken (New Jersey): John Wiley and Sons. pp 17–41.
11. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24: 332–340.
12. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22: 225–231.
13. Philippe H, Delsuc F, Brinkmann H, Lartillot N (2005) Phylogenomics. Annu Rev Ecol Evol Syst 36: 541–562.
14. Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62: 2080–2091.
15. Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24: 539–551.
16. Felsenstein J (2004) Inferring phylogenies. Sunderland (Massachusetts): Sinauer Associates. 645 p.
17. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3: e123. doi:10.1371/journal.pcbi.0030123.
18. Fitch WM (2000) Homology: a personal view on some of the problems. Trends Genet 16: 227–231.
19. van Dongen S (2000) Graph clustering by flow simulation [PhD dissertation]. Utrecht, The Netherlands: University of Utrecht, Available: http://igitur-archive.library.uu.nl/dissertations/1895620/inhoud.htm. Accessed 9 February 2011.
20. Li L, Stoeckert CJ Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

21. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22–28.
22. Schreiber F, Pick K, Erpenbeck D, Worheide G, Morgenstern B (2009) OrthoSelect: a protocol for selecting orthologous groups in phylogenomics. BMC Bioinformatics 10: 219.
23. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52: 540–542.
24. Haen KM, Lang BF, Pomponi SA, Lavrov DV (2007) Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? Mol Biol Evol 24: 1518–1527.
25. Kobayashi M, Wada H, Satoh N (1996) Early evolution of the Metazoa and phylogenetic status of diploblasts as inferred from amino acid sequence of elongation factor-1 alpha. Mol Phylogenet Evol 5: 414–422.
26. Medina M, Collins AG, Silberman JD, Sogin ML (2001) Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. Proc Natl Acad Sci U S A 98: 9707–9712.
27. Rokas A, King N, Finnerty J, Carroll SB (2003) Conflicting phylogenetic signals at the base of the metazoan tree. Evol Dev 5: 346–359.
28. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol 26: 2261–2274.
29. Galtier N (2007) A model of horizontal gene transfer and the bacterial phylogeny problem. Syst Biol 56: 633–642.
30. Brochier C, Bapteste E, Moreira D, Philippe H (2002) Eubacterial phylogeny based on translational apparatus proteins. Trends Genet 18: 1–5.
31. Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, et al. (2006) Mitochondrial genome of Trichoplax adhaerens supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A 103: 8751–8756.
32. Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27: 1983–1987.
33. Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38: 297–309.
34. Baurain D, Brinkmann H, Philippe H (2007) Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol 24: 6–9.
35. Hillis DM (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 47: 3–8.
36. Wiens JJ (2005) Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol 54: 731–742.
37. Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51: 588–598.
38. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol 58: 130–145.
39. Hartmann S, Vision TJ (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol 8: 95.
40. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, et al. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21: 1740–1752.
41. Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52: 528–538.
42. Wiens JJ, Moen DS (2008) Missing data and the accuracy of Bayesian phylogenetics. J Syst Evol 46: 307–314.
43. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, ed. Mammalian protein metabolism. New York: Academic Press. pp 21–132.
44. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
45. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21: 1095–1109.
46. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci 276: 4261–4270.
47. Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, et al. (1994) Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J Evol Biol 7: 247–265.
48. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, et al. (2010) A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol 27: 2541–2464.
49. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361–375.
50. Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42: 587–596.
51. Kupczok A, Schmidt HA, von Haeseler A (2010) Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol 5: 37.
52. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5: e1000392. doi:10.1371/journal.pcbi.1000392.
53. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.
54. Roure B, Rodriguez-Ezpeleta N, Philippe H (2007) SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol 7(Suppl 1): S2.
55. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, et al. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci U S A 99: 1414–1419.
56. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20: 1692–1704.
57. Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol 23: 1762–1775.
58. Redelings BD, Suchard MA (2005) Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54: 401–418.
59. Lunter G, Miklos I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6: 83.
60. Lartillot N (2006) Conjugate Gibbs sampling for Bayesian phylogenetic models. J Comput Biol 13: 43–63.
61. Stamatakis A, Ott M (2008) Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos Trans R Soc Lond B Biol Sci 363: 3977–3984.
62. de Koning AP, Gu W, Pollock DD (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. Mol Biol Evol 27: 249–265.
63. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368–376.
64. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
65. Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of protein sequence and structure. Washington (District of Columbia): National Biomedical Research Foundation. pp 89–99.
66. Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92: 11317–11321.
67. Yang Z, Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12: 451–458.
68. Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11: 367–370.
69. Kolaczkowski B, Thornton JW (2008) A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol 25: 1054–1066.
70. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, et al. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56: 389–399.
71. Nishihara H, Okada N, Hasegawa M (2007) Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol 8: R199.
72. Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25: 842–858.
73. Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. PLoS Comput Biol 5: e1000501. doi:10.1371/journal.pcbi.1000501.
74. Houliston E, Momose T, Manuel M (2010) Clytia hemisphaerica: a jellyfish cousin joins the laboratory. Trends Genet 26: 159–167.
75. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.
76. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286–2288.

nature microbiology

LETTERS

PUBLISHED: 11 APRIL 2016 | ARTICLE NUMBER: 16048 | DOI: 10.1038/NMICROBIOL.2016.48

OPEN

A new view of the tree of life

Laura A. Hug[1†], Brett J. Baker[2], Karthik Anantharaman[1], Christopher T. Brown[3], Alexander J. Probst[1], Cindy J. Castelle[1], Cristina N. Butterfield[1], Alex W. Hernsdorf[3], Yuki Amano[4], Kotaro Ise[4], Yohey Suzuki[5], Natasha Dudek[6], David A. Relman[7,8], Kari M. Finstad[9], Ronald Amundson[9], Brian C. Thomas[1] and Jillian F. Banfield[1,9*]

The tree of life is one of the most important organizing principles in biology[1]. Gene surveys suggest the existence of an enormous number of branches[2], but even an approximation of the full scale of the tree has remained elusive. Recent depictions of the tree of life have focused either on the nature of deep evolutionary relationships[3–5] or on the known, well-classified diversity of life with an emphasis on eukaryotes[6]. These approaches overlook the dramatic change in our understanding of life's diversity resulting from genomic sampling of previously unexamined environments. New methods to generate genome sequences illuminate the identity of organisms and their metabolic capacities, placing them in community and ecosystem contexts[7,8]. Here, we use new genomic data from over 1,000 uncultivated and little known organisms, together with published sequences, to infer a dramatically expanded version of the tree of life, with Bacteria, Archaea and Eukarya included. The depiction is both a global overview and a snapshot of the diversity within each major lineage. The results reveal the dominance of bacterial diversification and underline the importance of organisms lacking isolated representatives, with substantial evolution concentrated in a major radiation of such organisms. This tree highlights major lineages currently underrepresented in biogeochemical models and identifies radiations that are probably important for future evolutionary analyses.

Early approaches to describe the tree of life distinguished organisms based on their physical characteristics and metabolic features. Molecular methods dramatically broadened the diversity that could be included in the tree because they circumvented the need for direct observation and experimentation by relying on sequenced genes as markers for lineages. Gene surveys, typically using the small subunit ribosomal RNA (SSU rRNA) gene, provided a remarkable and novel view of the biological world[1,9,10], but questions about the structure and extent of diversity remain. Organisms from novel lineages have eluded surveys, because many are invisible to these methods due to sequence divergence relative to the primers commonly used for gene amplification[7,11]. Furthermore, unusual sequences, including those with unexpected insertions, may be discarded as artefacts[7].

Whole genome reconstruction was first accomplished in 1995 (ref. 12), with a near-exponential increase in the number of draft genomes reported each subsequent year. There are 30,437 genomes from all three domains of life—Bacteria, Archaea and Eukarya—which are currently available in the Joint Genome Institute's Integrated Microbial Genomes database (accessed 24 September 2015). Contributing to this expansion in genome numbers are single cell genomics[13] and metagenomics studies. Metagenomics is a shotgun sequencing-based method in which DNA isolated directly from the environment is sequenced, and the reconstructed genome fragments are assigned to draft genomes[14]. New bioinformatics methods yield complete and near-complete genome sequences, without a reliance on cultivation or reference genomes[7,15]. These genome- (rather than gene) based approaches provide information about metabolic potential and a variety of phylogenetically informative sequences that can be used to classify organisms[16]. Here, we have constructed a tree of life by making use of genomes from public databases and 1,011 newly reconstructed genomes that we recovered from a variety of environments (see Methods).

To render this tree of life, we aligned and concatenated a set of 16 ribosomal protein sequences from each organism. This approach yields a higher-resolution tree than is obtained from a single gene, such as the widely used 16S rRNA gene[16]. The use of ribosomal proteins avoids artefacts that would arise from phylogenies constructed using genes with unrelated functions and subject to different evolutionary processes. Another important advantage of the chosen ribosomal proteins is that they tend to be syntenic and co-located in a small genomic region in Bacteria and Archaea, reducing binning errors that could substantially perturb the geometry of the tree. Included in this tree is one representative per genus for all genera for which high-quality draft and complete genomes exist (3,083 organisms in total).

Despite the methodological challenges, we have included representatives of all three domains of life. Our primary focus relates to the status of Bacteria and Archaea, as these organisms have been most difficult to profile using macroscopic approaches, and substantial progress has been made recently with acquisition of new genome sequences[7,8,13]. The placement of Eukarya relative to Bacteria and Archaea is controversial[1,4,5,17,18]. Eukaryotes are believed to be evolutionary chimaeras that arose via endosymbiotic fusion, probably involving bacterial and archaeal cells[19]. Here, we do not attempt to confidently resolve the placement of the Eukarya. We position them using sequences of a subset of their nuclear-encoded ribosomal proteins, an approach that classifies them based on the inheritance of their information systems as opposed to lipid or other cellular structures[5].

Figure 1 presents a new view of the tree of life. This is one of a relatively small number of three-domain trees constructed from molecular information so far, and the first comprehensive tree to

[1]Department of Earth and Planetary Science, UC Berkeley, Berkeley, California 94720, USA. [2]Department of Marine Science, University of Texas Austin, Port Aransas, Texas 78373, USA. [3]Department of Plant and Microbial Biology, UC Berkeley, Berkeley, California 94720, USA. [4]Sector of Decommissioning and Radioactive Wastes Management, Japan Atomic Energy Agency, Ibaraki 319-1184, Japan. [5]Graduate School of Science, The University of Tokyo, Tokyo 113-8654, Japan. [6]Department of Ecology and Evolutionary Biology, UC Santa Cruz, Santa Cruz, California 95064, USA. [7]Departments of Medicine and of Microbiology and Immunology, Stanford University, Stanford, California 94305, USA. [8]Veterans Affairs Palo Alto Health Care System, Palo Alto, California 94304, USA. [9]Department of Environmental Science, Policy, and Management, UC Berkeley, Berkeley, California 94720, USA. [†]Present address: Department of Biology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. *e-mail: jbanfield@berkeley.edu

© 2016 Macmillan Publishers Limited. All rights reserved

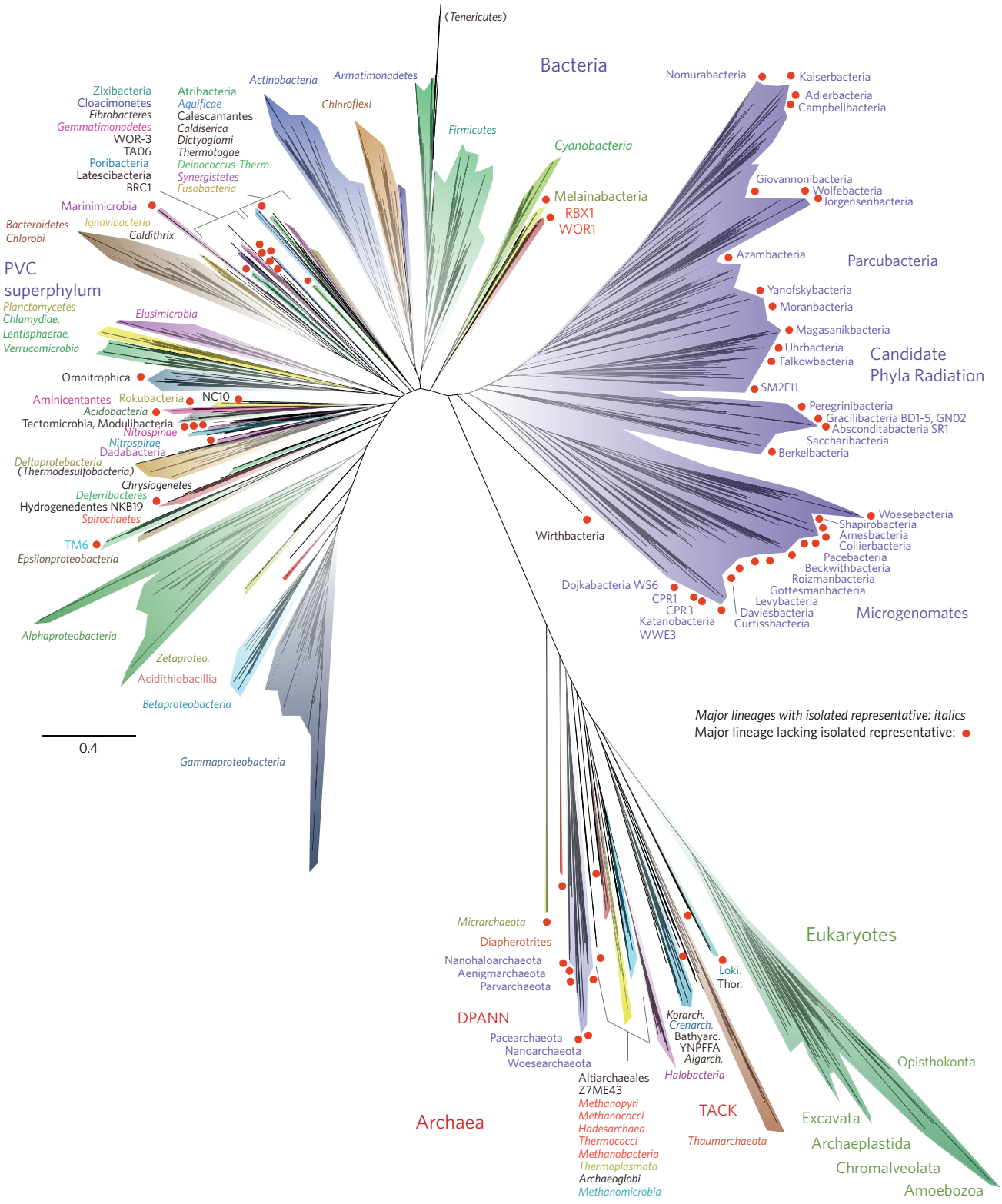


Figure 1 | A current view of the tree of life, encompassing the total diversity represented by sequenced genomes. The tree includes 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups. Major lineages are assigned arbitrary colours and named, with well-characterized lineage names, in italics. Lineages lacking an isolated representative are highlighted with non-italicized names and red dots. For details on taxon sampling and tree inference, see Methods. The names Tenericutes and Thermodesulfobacteria are bracketed to indicate that these lineages branch within the Firmicutes and the Deltaproteobacteria, respectively. Eukaryotic supergroups are noted, but not otherwise delineated due to the low resolution of these lineages. The CPR phyla are assigned a single colour as they are composed entirely of organisms without isolated representatives, and are still in the process of definition at lower taxonomic levels. The complete ribosomal protein tree is available in rectangular format with full bootstrap values as Supplementary Fig. 1 and in Newick format in Supplementary Dataset 2.

© 2016 Macmillan Publishers Limited. All rights reserved

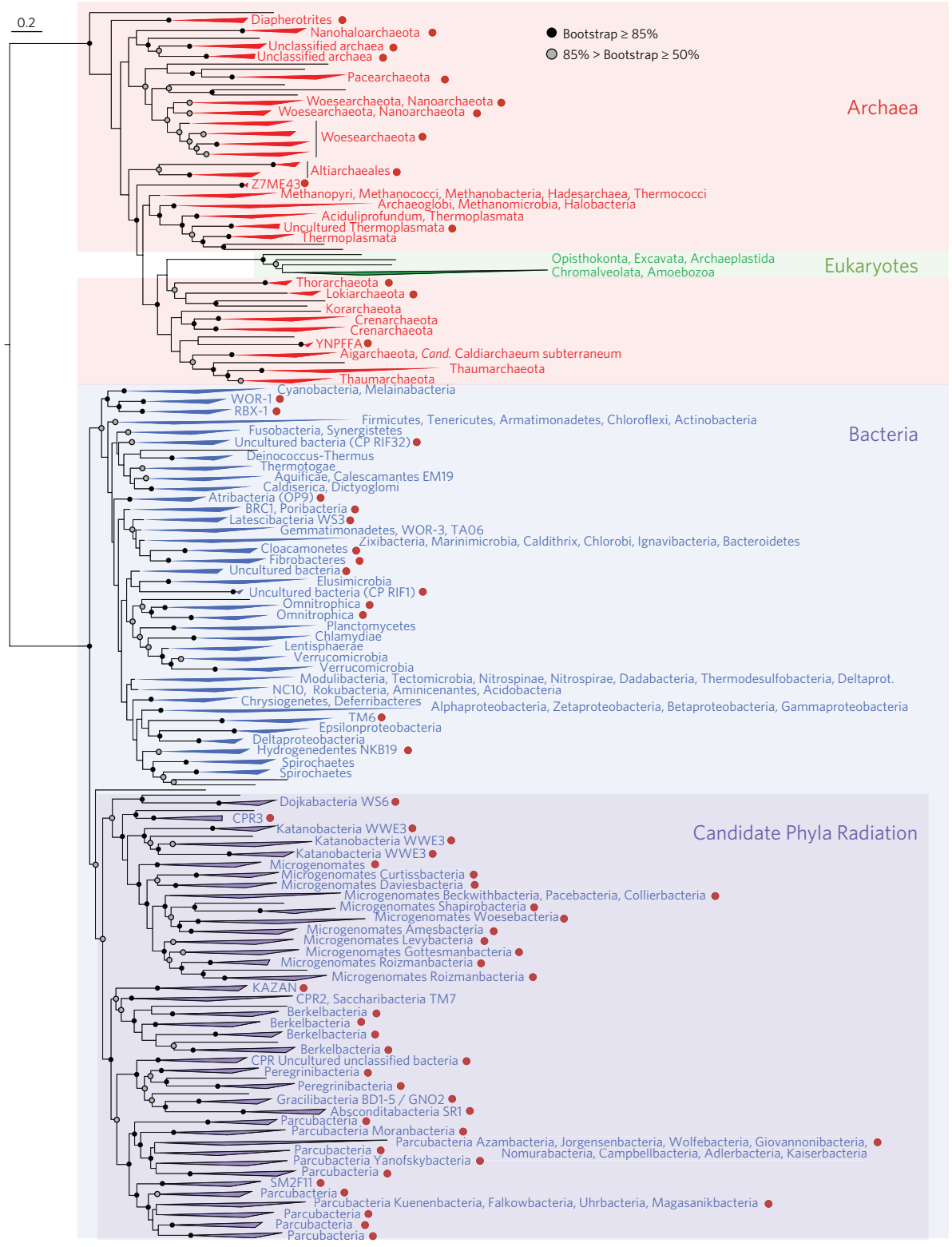


Figure 2 | A reformatted view of the tree in Fig. 1 in which each major lineage represents the same amount of evolutionary distance. The threshold for groups (coloured wedges) was an average branch length of <0.65 substitutions per site. Notably, some well-accepted phyla become single groups and others are split into multiple distinct groups. We undertook this analysis to provide perspective on the structure of the tree, and do not propose the resulting groups to have special taxonomic status. The massive scale of diversity in the CPR and the large fraction of major lineages that lack isolated representatives (red dots) are apparent from this analysis. Bootstrap support values are indicated by circles on nodes—black for support of 85% and above, grey for support from 50 to 84%. The complete ribosomal protein tree is available in rectangular format with full bootstrap values as Supplementary Fig. 1 and in Newick format in Supplementary Dataset 2.

© 2016 Macmillan Publishers Limited. All rights reserved

be published since the development of genome-resolved metagenomics. We highlight all major lineages with genomic representation, most of which are phylum-level branches (see Supplementary Fig. 1 for full bootstrap support values). However, we separately identify the Classes of the Proteobacteria, because the phylum is not monophyletic (for example, the Deltaproteobacteria branch away from the other Proteobacteria, as previously reported[2,20]).

The tree in Fig. 1 recapitulates expected organism groupings at most taxonomic levels and is largely congruent with the tree calculated using traditional SSU rRNA gene sequence information (Supplementary Fig. 2). The support values for taxonomic groups are strong at the Species through Class levels (>85%), with moderate-to-strong support for Phyla (>75% in most cases), but the branching order of the deepest branches cannot be confidently resolved (Supplementary Fig. 1). The lower support for deep branch placements is a consequence of our prioritization of taxon sampling over number of genes used for tree construction. As proposed recently, the Eukarya, a group that includes protists, fungi, plants and animals, branches within the Archaea, specifically within the TACK superphylum[21] and sibling to the Lokiarchaeota[22]. Interestingly, this placement is not evident in the SSU rRNA tree, which has the three-domain topology proposed by Woese and co-workers in 1990[1] (Supplementary Fig. 2). The two-domain Eocyte tree and the three-domain tree are competing hypotheses for the origin of Eukarya[5]; further analyses to resolve these and other deep relationships will be strengthened with the availability of genomes for a greater diversity of organisms. Important advantages of the ribosomal protein tree compared with the SSU rRNA gene tree are that it includes organisms with incomplete or unavailable SSU rRNA gene sequences and more strongly resolves the deeper radiations. Ribosomal proteins have been shown to contain compositional biases across the three domains, driven by thermophilic, mesophilic and halophilic lifestyles as well as by a primitive genetic code[23]. Continued expansion of the number of genome sequences for non-extremophile Archaea, such as the DPANN lineages[8,13], may allow clarification of these compositional biases.

A striking feature of this tree is the large number of major lineages without isolated representatives (red dots in Fig. 1). Many of these lineages are clustered together into discrete regions of the tree. Of particular note is the Candidate Phyla Radiation (CPR)[7], highlighted in purple in Fig. 1. Based on information available from hundreds of genomes from genome-resolved metagenomics and single-cell genomics methods to date, all members have relatively small genomes and most have somewhat (if not highly) restricted metabolic capacities[7,13,24]. Many are inferred (and some have been shown) to be symbionts[7,25,26]. Thus far, all cells lack complete citric acid cycles and respiratory chains and most have limited or no ability to synthesize nucleotides and amino acids. It remains unclear whether these reduced metabolisms are a consequence of superphylum-wide loss of capacities or if these are inherited characteristics that hint at an early metabolic platform for life. If inherited, then adoption of symbiotic lifestyles may have been a later innovation by these organisms once more complex organisms appeared.

Figure 2 presents another perspective, where the major lineages of the tree are defined using evolutionary distance, so that the main groups become apparent without bias arising from historical naming conventions. This depiction uses the same inferred tree as in Fig. 1, but with groups defined on the basis of average branch length to the leaf taxa. We chose an average branch length that best recapitulated the current taxonomy (smaller values fragmented many currently accepted phyla and larger values collapsed accepted phyla into very few lineages, see Methods). Evident in Fig. 2 is the enormous extent of evolution that has occurred within the CPR. The diversity within the CPR could be a result of the early emergence of this group and/or a consequence of rapid evolution related to symbiotic lifestyles. The CPR is early-emerging on the ribosomal protein tree (Fig. 1), but not in the SSU rRNA tree (Supplementary Fig. 2). Regardless of branching order, the CPR, in combination with other lineages that lack isolated representatives (red dots in Fig. 2), clearly comprises the majority of life's current diversity.

Domain Bacteria includes more major lineages of organisms than the other Domains. We do not attribute the smaller scope of the Archaea relative to Bacteria to sampling bias because metagenomics and single-cell genomics methods detect members of both domains equally well. Consistent with this view, Archaea are less prominent and less diverse in many ecosystems (for example, seawater[27], hydrothermal vents[28], the terrestrial subsurface[15] and human-associated microbiomes[29]). The lower apparent phylogenetic diversity of Eukarya is fully expected, based on their comparatively recent evolution.

The tree of life as we know it has dramatically expanded due to new genomic sampling of previously enigmatic or unknown microbial lineages. This depiction of the tree captures the current genomic sampling of life, illustrating the progress that has been made in the last two decades following the first published genome. What emerges from analysis of this tree is the depth of evolutionary history that is contained within the Bacteria, in part due to the CPR, which appears to subdivide the domain. Most importantly, the analysis highlights the large fraction of diversity that is currently only accessible via cultivation-independent genome-resolved approaches.

Methods

A data set comprehensively covering the three domains of life was generated using publicly available genomes from the Joint Genome Institute's IMG-M database (img.jgi.doe.gov), a previously developed data set of eukaryotic genome information[30], previously published genomes derived from metagenomic data sets[7,8,31,32] and newly reconstructed genomes from current metagenome projects (see Supplementary Table 1 for NCBI accession numbers). From IMG-M, genomes were sampled such that a single representative for each defined genus was selected. For phyla and candidate phyla lacking full taxonomic definition, every member of the phylum was initially included. Subsequently, these radiations were sampled to an approximate genus level of divergence based on comparison with taxonomically described phyla, thus removing strain- and species-level overlaps. Finally, initial tree reconstructions identified aberrant long-branch attraction effects placing the Microsporidia, a group of parasitic fungi, with the Korarchaeota. The Microsporidia are known to contribute long branch attraction artefacts confounding placement of the Eukarya[33], and were subsequently removed from the analysis.

This study includes 1,011 organisms from lineages for which genomes were not previously available. The organisms were present in samples collected from a shallow aquifer system, a deep subsurface research site in Japan, a salt crust in the Atacama Desert, grassland meadow soil in northern California, a CO_2-rich geyser system, and two dolphin mouths. Genomes were reconstructed from metagenomes as described previously[7]. Genomes were only included if they were estimated to be >70% complete based on presence/absence of a suite of 51 single copy genes for Bacteria and 38 single copy genes for Archaea. Genomes were additionally required to have consistent nucleotide composition and coverage across scaffolds, as determined using the ggkbase binning software (ggkbase.berkeley.edu), and to show consistent placement across both SSU rRNA and concatenated ribosomal protein phylogenies. This contributed marker gene information for 1,011 newly sampled organisms, whose genomes were reconstructed for metabolic analyses to be published separately.

The concatenated ribosomal protein alignment was constructed as described previously[16]. In brief, the 16 ribosomal protein data sets (ribosomal proteins L2, L3, L4, L5, L6, L14, L16, L18, L22, L24, S3, S8, S10, S17 and S19) were aligned independently using MUSCLE v. 3.8.31 (ref. 34). Alignments were trimmed to remove ambiguously aligned C and N termini as well as columns composed of more than 95% gaps. Taxa were removed if their available sequence data represented less than 50% of the expected alignment columns (90% of taxa had more than 80% of the expected alignment columns). The 16 alignments were concatenated, forming a final alignment comprising 3,083 genomes and 2,596 amino-acid positions. A maximum likelihood tree was constructed using RAxML v. 8.1.24 (ref. 35), as implemented on the CIPRES web server[36], under the LG plus gamma model of evolution (PROTGAMMALG in the RAxML model section), and with the number of bootstraps automatically determined (MRE-based bootstopping criterion). A total of 156 bootstrap replicates were conducted under the rapid bootstrapping algorithm, with 100 sampled to generate proportional support values. The full tree inference required 3,840 computational hours on the CIPRES supercomputer.

© 2016 Macmillan Publishers Limited. All rights reserved

To construct Fig. 2, we collapsed branches based on an average branch length criterion. Average branch length calculations were implemented in the Interactive Tree of Life online interface[37] using the formula:

Average branch length=mean([root distance to tip]–[root distance to node]) for all tips connecting to a node.

We tested values between 0.25 and 0.75 at 0.05 intervals, and selected a final threshold of <0.65 based on generation of a similar number of major lineages as compared to the taxonomy-guided clustering view in Fig. 1. The taxonomy view identified 26 archaeal and 74 bacterial phylum-level lineages (counting the Microgenomates and Parcubacteria as single phyla each), whereas an average branch length of <0.65 resulted in 28 archaeal and 76 bacterial clades.

For a companion SSU rRNA tree, an alignment was generated from all SSU rRNA genes available from the genomes of the organisms included in the ribosomal protein data set. For organisms with multiple SSU rRNA genes, one representative gene was kept for the analysis, selected randomly. As genome sampling was confined to the genus level, we do not anticipate this selection process will have any impact on the resultant tree. All SSU rRNA genes longer than 600 bp were aligned using the SINA alignment algorithm through the SILVA web interface[38,39]. The full alignment was stripped of columns containing 95% or more gaps, generating a final alignment containing 1,871 taxa and 1,947 alignment positions. A maximum likelihood tree was inferred as described for the concatenated ribosomal protein trees, with RAxML run using the GTRCAT model of evolution. The RAxML inference included the calculation of 300 bootstrap iterations (extended majority rules-based bootstopping criterion), with 100 randomly sampled to determine support values.

To test the effect of site selection stringency on the inferred phylogenies, we stripped the alignments of columns containing up to 50% gaps (compared with the original trimming of 95% gaps). For the ribosomal protein alignment, this resulted in a 14% reduction in alignment length (to 2,232 positions) and a 44.6% reduction in computational time (~2,100 h). For the SSU rRNA gene alignment, stripping columns with 50% or greater gaps reduced the alignment by 24% (to 1,489 positions) and the computation time by 28%. In both cases, the topology of the tree with the best likelihood was not changed significantly. The ribosomal protein resolved a two-domain tree with the Eukarya sibling to the Lokiarcheaota, while the SSU rRNA tree depicts a three-domain tree. The position of the CPR as deep-branching on the ribosomal protein tree and within the Bacteria on the SSU rRNA tree was also consistent. The alignments and inferred trees under the more stringent gap stripping are available upon request.

Nomenclature. We have included names for two lineages for which we have previously published complete genomes[40]. At the time of submission of the paper describing these genomes[40], the reviewer community was not uniformly open to naming lineages of uncultivated organisms based on such information. Given that this practice is now widely used, we re-propose the names for these phyla. Specifically, for WWE3 we suggest the name Katanobacteria from the Hebrew 'katan', which means 'small', and for SR1 we suggest the name Absconditabacteria from the Latin 'Abscondo' meaning 'hidden', as in 'shrouded'.

Accession codes. NCBI and/or JGI IMG accession numbers for all genomes used in this study are listed in Supplementary Table 1. Additional ribosomal protein gene and 16S rRNA gene sequences used in this study have been deposited in Genbank under accession numbers KU868081–KU869521. The concatenated ribosomal protein and SSU rRNA alignments used for tree reconstruction are included as separate files in the Supplementary Information.

Received 25 January 2016; accepted 10 March 2016; published 11 April 2016

References

1. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87,** 4576–4579 (1990).
2. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Rev. Microbiol.* **12,** 635–645 (2014).
3. Gouy, R., Baurain, D. & Philippe, H. Rooting the tree of life: the phylogenetic jury is still out. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20140329 (2015).
4. Pace, N. R. Time for a change. *Nature* **441,** 289 (2006).
5. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
6. Hinchliff, C. E. *et al.* Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl Acad. Sci. USA* **112,** 12764–12769 (2015).
7. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).
8. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25,** 690–701 (2015).
9. Pace, N. R. Mapping the tree of life: progress and prospects. *Microbiol. Mol. Biol. Rev.* **73,** 565–576 (2009).
10. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA* **82,** 6955–6959 (1985).
11. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41,** e1 (2013).
12. Fleischmann, R. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
13. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
14. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10,** R85 (2009).
15. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Commun.* **4,** 2120 (2013).
16. Hug, L. A. *et al.* Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1,** 22 (2013).
17. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaebacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105,** 20356–20361 (2008).
18. Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl Acad. Sci. USA* **81,** 3786–3790 (1984).
19. Martin, W. F., Garg, S. & Zimorski, V. Endosymbiotic theories for eukaryote origin. *Phil. Trans. R. Soc. B Biol. Sci.* **370,** 20140330 (2015).
20. Garrity, G., Boone, D. R. & Castenholtz, R. W. (eds) *Bergey's Manual of Systematic Bacteriology. Volume One: The Archaea and the Deeply Branching and Phototrophic Bacteria* 2nd edn (Springer, 2001).
21. Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19,** 580–587 (2011).
22. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179 (2015).
23. Fournier, G. P. & Gogarten, J. P. Rooting the ribosomal tree of life. *Mol. Biol. Evol.* **27,** 1792–1801 (2010).
24. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337,** 1661–1665 (2012).
25. He, X. *et al.* Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl Acad. Sci. USA* **112,** 244–249 (2015).
26. Nelson, W. C. & Stegen, J. C. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front. Microbiol.* **6,** 713 (2015).
27. Lloyd, K. G., May, M. K., Kevorkian, R. T. & Steen, A. D. Meta-analysis of quantification methods shows that archaea and bacteria have similar abundances in the subseafloor. *Appl. Environ. Microbiol.* **79,** 7790–7799 (2013).
28. Sievert, S. M., Ziebis, W., Kuever, J. & Sahm, K. Relative abundance of Archaea and Bacteria along a thermal gradient of a shallow-water hydrothermal vent quantified by rRNA slot-blot hybridization. *Microbiology* **146,** 1287–1293 (2000).
29. Hoffmann, C. *et al.* Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. *PLoS ONE* **8,** e66019 (2013).
30. Brown, M. W. *et al.* Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc. Biol. Sci.* **280,** 20131755 (2013).
31. Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* **18,** 159–173 (2016).
32. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3,** 14 (2015).
33. Inagaki, Y., Susko, E., Fast, N. M. & Roger, A. J. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. *Mol. Biol. Evol.* **21,** 1340–1349 (2004).
34. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
35. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
36. Miller, M., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gatew. Comput. Environ. Work* 1–8 (2010).
37. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39,** W475–W478 (2011).
38. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28,** 1823–1829 (2012).
39. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).

© 2016 Macmillan Publishers Limited. All rights reserved

40. Kantor, R. S. *et al.* Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *MBio* **4,** e00708–13 (2013).

Acknowledgements
This research was largely supported by the Lawrence Berkeley National Laboratory (LBNL) Genomes to Watershed Scientific Focus Area funded by the US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research (OBER) under contract no. DE-AC02-05CH11231. Additional support was provided by LBNL EFRC award no. DE-AC02-05CH11231, NASA NESSF grant no. 12-PLANET12R-0025 and NSF DEB grant no. 1406956, DOE OBER grant no. DOE-SC10010566, Office of Naval Research grants nos. N00014-07-1-0287, N00014-10-1-0233 and N00014-11-1-0918, and by the Thomas C. and Joan M. Merigan Endowment at Stanford University. In addition, funding was provided by the Ministry of Economy, Trade and Industry of Japan. The authors thank J. Eisen for comments, S. Venn-Watson, K. Carlin and E. Jensen (US Navy Marine Mammal Program) for dolphin samples, K.W. Seitz for sequence submission assistance, and the DOE Joint Genome Institute for generating the metagenome sequence via the Community Science Program.

Author contributions
L.A.H. and J.F.B. defined the research objective. L.A.H. generated data sets and conducted phylogenetic tree inferences. L.A.H., B.J.B. and J.F.B. conducted data analysis. L.A.H., B.J.B., K.A., C.T.B., A.J.P., C.J.C., C.N.B., A.W.H., Y.A., K.I., Y.S., N.D., D.A.R., K.M.F., R.A., B.C.T. and J.F.B. contributed to metagenome binning and genome analysis. L.A.H. and J.F.B. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Additional information
Supplementary information is available online. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.F.B.

Competing interests
The authors declare no competing financial interests.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© 2016 Macmillan Publishers Limited. All rights reserved

A.5 Genomics

Recommended reading for Chapter 5 (listed in Section 5.10).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Goodwin et al. (2016) — ten years of next-generation sequencing; how the technologies work and what their errors are.
  DOI · via SU library (SU login)

  Coming of age: ten years of next-generation sequencing technologies. Goodwin et al. (2016). *Nature Reviews Genetics*. doi:10.1038/nrg.2016.49

  > Since the completion of the human genome project in 2003, extraordinary progress has been made in genome sequencing technologies, which has led to a decreased cost per megabase and an increase in the number and diversity of sequenced genomes. An astonishing complexity of genome architecture has been revealed, bringing these sequencing technologies to even greater advancements. Some approaches maximize the number of bases sequenced in the least amount of time, generating a wealth of data that can be used to understand increasingly complex phenotypes. Alternatively, other approaches now aim to sequence longer contiguous pieces of DNA, which are essential for resolving structurally complex regions. These and other strategies are providing researchers and clinicians a variety of tools to probe genomes in greater depth, leading to an enhanced understanding of how genome sequence variants underlie phenotype and disease.

- Compeau et al. (2011) — how to apply de Bruijn graphs to genome assembly; the key algorithmic idea, very clearly explained.
  DOI · via SU library (SU login)

  How to apply de Bruijn graphs to genome assembly. Compeau et al. (2011). *Nature Biotechnology*. doi:10.1038/nbt.2023

- Langmead et al. (2009) — Bowtie; read mapping with the Burrows–Wheeler transform.
  full text below · DOI · open access (CC BY 2.0)

  Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Langmead et al. (2009). *Genome Biology*. doi:10.1186/gb-2009-10-3-r25

  > Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source (http://bowtie.cbcb.umd.edu).

- Nielsen et al. (2011) — genotype and SNP calling from sequencing data; the probabilistic framework.
  DOI · via SU library (SU login)

  Genotype and SNP calling from next-generation sequencing data. Nielsen et al. (2011). *Nature Reviews Genetics*. doi:10.1038/nrg2986

  > Meaningful analysis of next-generation sequencing (NGS) data, which are produced extensively by genetics and genomics studies, relies crucially on the accurate calling of SNPs and genotypes. Recently developed statistical methods both improve and quantify the considerable uncertainty associated with genotype calling, and will especially benefit the growing number of studies using low- to medium-coverage data. We review these methods and provide a guide for their use in NGS studies.

- Yandell and Ence (2012) — a beginner's guide to eukaryotic genome annotation.
  DOI · via SU library (SU login)

  A beginner's guide to eukaryotic genome annotation. Yandell & Ence (2012). *Nature Reviews Genetics*. doi:10.1038/nrg3174

  > The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.

- Quince et al. (2017) — shotgun metagenomics from sampling to analysis.
  DOI · via SU library (SU login)

  Shotgun metagenomics, from sampling to analysis. Quince et al. (2017). *Nature Biotechnology*. doi:10.1038/nbt.3935

  > Diverse microbial communities of bacteria, archaea, viruses and single-celled eukaryotes have crucial roles in the environment and in human health. However, microbes are frequently difficult to culture in the laboratory, which can confound cataloging of members and understanding of how communities function. High-throughput sequencing technologies and a suite of computational pipelines have been combined into shotgun metagenomics methods that have transformed microbiology. Still, computational approaches to overcome the challenges that affect both assembly-based and mapping-based metagenomic profiling, particularly of high-complexity samples or environments containing organisms with limited similarity to sequenced genomes, are needed. Understanding the functions and characterizing specific strains of these communities offers biotechnological promise in therapeutic discovery and innovative ways to synthesize products using microbial factories and can pinpoint the contributions of microorganisms to planetary, animal and human health.

The open-access papers are reproduced below under their original licences.

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome ***Biology*** 2009, **10:**R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

© 2009 Langmead *et al.*; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.

Rationale

Improvements in the efficiency of DNA sequencing have both broadened the applications for sequencing and dramatically increased the size of sequencing datasets. Technologies from Illumina (San Diego, CA, USA) and Applied Biosystems (Foster City, CA, USA) have been used to profile methylation patterns (MeDIP-Seq) [1], to map DNA-protein interactions (ChIP-Seq) [2], and to identify differentially expressed genes (RNA-Seq) [3] in the human genome and other species. The Illumina instrument was recently used to re-sequence three human genomes, one from a cancer patient and two from previously unsequenced ethnic groups [4-6]. Each of these studies required the alignment of large numbers of short DNA sequences ('short reads') onto the human genome. For example, two of the studies [4,5] used the short read alignment tool Maq [7] to align more than 130 billion bases (about 45× coverage) of short Illumina reads to a human reference genome in order to detect genetic variations. The third human re-sequencing study [6] used the SOAP program [8] to align more than 100 billion bases to the reference genome. In addition to these projects, the 1,000 Genomes project is in the process of using high-throughput sequencing instruments to sequence a total of about six trillion base pairs of human DNA [9].

With existing methods, the computational cost of aligning many short reads to a mammalian genome is very large. For example, extrapolating from the results presented here in Tables 1 and 2, one can see that Maq would require more than 5 central processing unit (CPU)-months and SOAP more than 3 CPU-years to align the 140 billion bases from the study by Ley and coworkers [5]. Although using Maq or SOAP for this purpose has been shown to be feasible by using multiple CPUs, there is a clear need for new tools that consume less time and computational resources.

Maq and SOAP take the same basic algorithmic approach as other recent read mapping tools such as RMAP [10], ZOOM [11], and SHRiMP [12]. Each tool builds a hash table of short oligomers present in either the reads (SHRiMP, Maq, RMAP, and ZOOM) or the reference (SOAP). Some employ recent theoretical advances to align reads quickly without sacrificing sensitivity. For example, ZOOM uses 'spaced seeds' to significantly outperform RMAP, which is based on a simpler algo-

Table 1

Bowtie alignment performance versus SOAP and Maq

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
Bowtie -v 2	Server	15 m 7 s	15 m 41 s	33.8	1,149	-	67.4
SOAP		91 h 57 m 35 s	91 h 47 m 46 s	0.10	13,619	351×	67.3
Bowtie	PC	16 m 41 s	17 m 57 s	29.5	1,353	-	71.9
Maq		17 h 46 m 35 s	17 h 53 m 7 s	0.49	804	59.8×	74.7
Bowtie	Server	17 m 58 s	18 m 26 s	28.8	1,353	-	71.9
Maq		32 h 56 m 53 s	32 h 58 m 39 s	0.27	804	107×	74.7

The performance and sensitivity of Bowtie v0.9.6, SOAP v1.10, and Maq v0.6.6 when aligning 8.84 M reads from the 1,000 Genome project (National Center for Biotechnology Information Short Read Archive: SRR001115) trimmed to 35 base pairs. The 'soap.contig' version of the SOAP binary was used. SOAP could not be run on the PC because SOAP's memory footprint exceeds the PC's physical memory. For the SOAP comparison, Bowtie was invoked with '-v 2' to mimic SOAP's default matching policy (which allows up to two mismatches in the alignment and disregards quality values). For the Maq comparison Bowtie is run with its default policy, which mimics Maq's default policy of allowing up to two mismatches during the first 28 bases and enforcing an overall limit of 70 on the sum of the quality values at all mismatched positions. To make Bowtie's memory footprint more comparable to Maq's, Bowtie is invoked with the '-z' option in all experiments to ensure only the forward or mirror index is resident in memory at one time. CPU, central processing unit.

rithm developed by Baeza-Yaetes and Perleberg [13]. Spaced seeds have been shown to yield higher sensitivity than contiguous seeds of the same length [14,15]. SHRiMP employs a combination of spaced seeds and the Smith-Waterman [16] algorithm to align reads with high sensitivity at the expense of speed. Eland is a commercial alignment program available from Illumina that uses a hash-based algorithm to align reads.

Bowtie uses a different and novel indexing strategy to create an ultrafast, memory-efficient short read aligner geared toward mammalian re-sequencing. In our experiments using reads from the 1,000 Genomes project, Bowtie aligns 35-base pair (bp) reads at a rate of more than 25 million reads per CPU-hour, which is more than 35 times faster than Maq and 300 times faster than SOAP under the same conditions (see Tables 1 and 2). Bowtie employs a Burrows-Wheeler index based on the full-text minute-space (FM) index, which has a memory footprint of only about 1.3 gigabytes (GB) for the human genome. The small footprint allows Bowtie to run on a typical desktop computer with 2 GB of RAM. The index is small enough to be distributed over the internet and to be stored on disk and re-used. Multiple processor cores can be used simultaneously to achieve even greater alignment speed. We have used Bowtie to align 14.3× coverage worth of human Illumina reads from the 1,000 Genomes project in about 14 hours on a single desktop computer with four processor cores.

Bowtie makes a number of compromises to achieve this speed, but these trade-offs are reasonable within the context of mammalian re-sequencing projects. If one or more exact matches exist for a read, then Bowtie is guaranteed to report one, but if the best match is an inexact one then Bowtie is not guaranteed in all cases to find the highest quality alignment. With its highest performance settings, Bowtie may fail to align a small number of reads with valid alignments, if those reads have multiple mismatches. If the stronger guarantees are desired, Bowtie supports options that increase accuracy at the cost of some performance. For instance, the '--best' option will guarantee that all alignments reported are best in terms

Table 2

Bowtie alignment performance versus Maq with filtered read set

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed up	Reads aligned (%)
Bowtie	PC	16 m 39 s	17 m 47 s	29.8	1,353	-	74.9
Maq		11 h 15 m 58 s	11 h 22 m 2 s	0.78	804	38.4×	78.0
Bowtie	Server	18 m 20 s	18 m 46 s	28.3	1,352	-	74.9
Maq		18 h 49 m 7 s	18 h 50 m 16 s	0.47	804	60.2×	78.0

Performance and sensitivity of Bowtie v0.9.6 and Maq v0.6.6 when the read set is filtered using Maq's 'catfilter' command to eliminate poly-A artifacts. The filter eliminates 438,145 out of 8,839,010 reads. Other experimental parameters are identical to those of the experiments in Table 1. CPU, central processing unit.

of minimizing mismatches in the seed portion of the read, although this option incurs additional computational cost.

With its default options, Bowtie's sensitivity measured in terms of reads aligned is equal to SOAP's and somewhat less than Maq's. Command line options allow the user to increase sensitivity at the cost of greater running time, and to enable Bowtie to report multiple hits for a read. Bowtie can align reads as short as four bases and as long as 1,024 bases. The input to a single run of Bowtie may comprise a mixture of reads with different lengths.

Bowtie description and results

Bowtie indexes the reference genome using a scheme based on the Burrows-Wheeler transform (BWT) [17] and the FM index [18,19]. A Bowtie index for the human genome fits in 2.2 GB on disk and has a memory footprint of as little as 1.3 GB at alignment time, allowing it to be queried on a workstation with under 2 GB of RAM.

The common method for searching in an FM index is the exact-matching algorithm of Ferragina and Manzini [18]. Bowtie does not simply adopt this algorithm because exact matching does not allow for sequencing errors or genetic variations. We introduce two novel extensions that make the technique applicable to short read alignment: a quality-aware backtracking algorithm that allows mismatches and favors high-quality alignments; and 'double indexing', a strategy to avoid excessive backtracking. The Bowtie aligner follows a policy similar to Maq's, in that it allows a small number of mismatches within the high-quality end of each read, and it places an upper limit on the sum of the quality values at mismatched alignment positions.

Burrows-Wheeler indexing

The BWT is a reversible permutation of the characters in a text. Although originally developed within the context of data compression, BWT-based indexing allows large texts to be searched efficiently in a small memory footprint. It has been applied to bioinformatics applications, including oligomer counting [20], whole-genome alignment [21], tiling microarray probe design [22], and Smith-Waterman alignment to a human-sized reference [23].

The Burrows-Wheeler transformation of a text T, BWT(T), is constructed as follows. The character $ is appended to T, where $ is not in T and is lexicographically less than all characters in T. The Burrows-Wheeler matrix of T is constructed as the matrix whose rows comprise all cyclic rotations of T$. The rows are then sorted lexicographically. BWT(T) is the sequence of characters in the rightmost column of the Burrows-Wheeler matrix (Figure 1a). BWT(T) has the same length as the original text T.

This matrix has a property called 'last first (LF) mapping'. The ith occurrence of character X in the last column corresponds to the same text character as the ith occurrence of X in the first

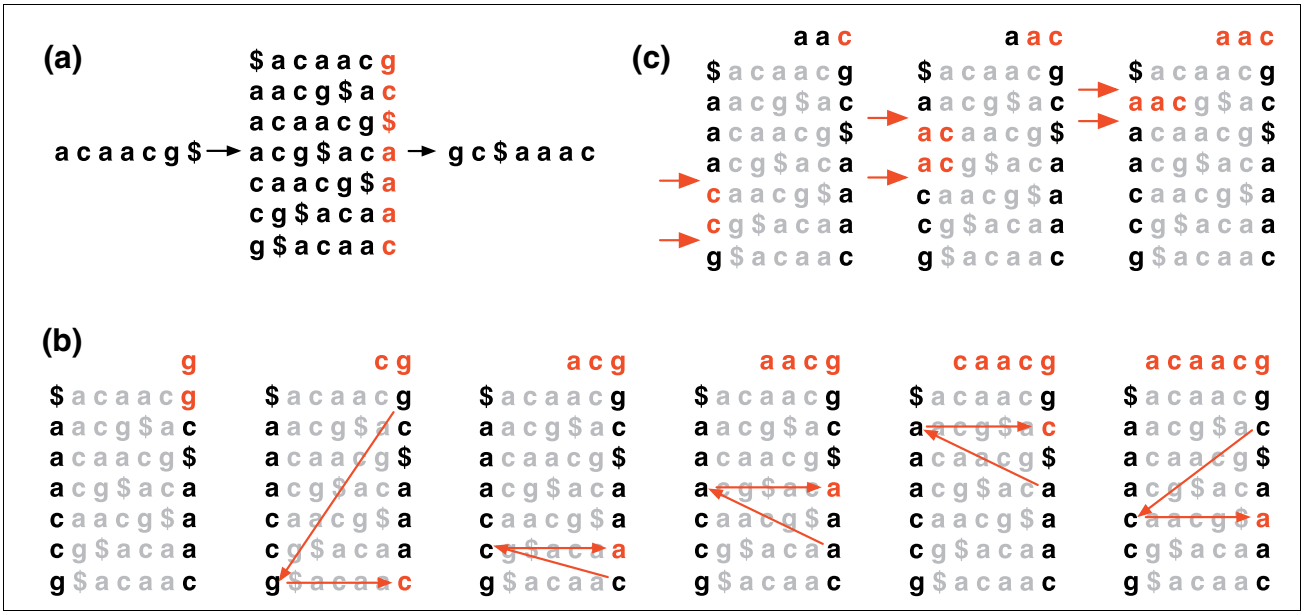


Figure 1
Burrows-Wheeler transform. **(a)** The Burrows-Wheeler matrix and transformation for 'acaacg'. **(b)** Steps taken by EXACTMATCH to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'. **(c)** UNPERMUTE repeatedly applies the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column).

column. This property underlies algorithms that use the BWT index to navigate or search the text. Figure 1b illustrates UNPERMUTE, an algorithm that applies the LF mapping repeatedly to re-create T from BWT(T).

The LF mapping is also used in exact matching. Because the matrix is sorted lexicographically, rows beginning with a given sequence appear consecutively. In a series of steps, the EXACTMATCH algorithm (Figure 1c) calculates the range of matrix rows beginning with successively longer suffixes of the query. At each step, the size of the range either shrinks or remains the same. When the algorithm completes, rows beginning with S_0 (the entire query) correspond to exact occurrences of the query in the text. If the range is empty, the text does not contain the query. UNPERMUTE is attributable to Burrows and Wheeler [17] and EXACTMATCH to Ferragina and Manzini [18]. See Additional data file 1 (Supplementary Discussion 1) for details.

Searching for inexact alignments

EXACTMATCH is insufficient for short read alignment because alignments may contain mismatches, which may be due to sequencing errors, genuine differences between reference and query organisms, or both. We introduce an alignment algorithm that conducts a backtracking search to quickly find alignments that satisfy a specified alignment policy. Each character in a read has a numeric quality value, with lower values indicating a higher likelihood of a sequencing error. Our alignment policy allows a limited number of mismatches and prefers alignments where the sum of the quality values at all mismatched positions is low.

The search proceeds similarly to EXACTMATCH, calculating matrix ranges for successively longer query suffixes. If the range becomes empty (a suffix does not occur in the text), then the algorithm may select an already-matched query position and substitute a different base there, introducing a mismatch into the alignment. The EXACTMATCH search resumes from just after the substituted position. The algorithm selects only those substitutions that are consistent with the alignment policy and which yield a modified suffix that occurs at least once in the text. If there are multiple candidate substitution positions, then the algorithm greedily selects a position with a minimal quality value.

Backtracking scenarios play out within the context of a stack structure that grows when a new substitution is introduced and shrinks when the aligner rejects all candidate alignments for the substitutions currently on the stack. See Figure 2 for an illustration of how the search might proceed.

In short, Bowtie conducts a quality-aware, greedy, randomized, depth-first search through the space of possible alignments. If a valid alignment exists, then Bowtie will find it (subject to the backtrack ceiling discussed in the following section). Because the search is greedy, the first valid alignment encountered by Bowtie will not necessarily be the 'best' in terms of number of mismatches or in terms of quality. The user may instruct Bowtie to continue searching until it can prove that any alignment it reports is 'best' in terms of number of mismatches (using the option --best). In our experience, this mode is two to three times slower than the default mode. We expect that the faster default mode will be preferred for large re-sequencing projects.

The user may also opt for Bowtie to report all alignments up to a specified number (option -k) or all alignments with no limit on the number (option -a) for a given read. If in the course of its search Bowtie finds N possible alignments for a given set of substitutions, but the user has requested only K alignments where K < N, Bowtie will report K of the N alignments selected at random. Note that these modes can be much slower than the default. In our experience, for example, -k 1 is more than twice as fast as -k 2.

Excessive backtracking

The aligner as described so far can, in some cases, encounter sequences that cause excessive backtracking. This occurs when the aligner spends most of its effort fruitlessly backtracking to positions close to the 3' end of the query. Bowtie mitigates excessive backtracking with the novel technique of 'double indexing'. Two indices of the genome are created: one containing the BWT of the genome, called the 'forward' index, and a second containing the BWT of the genome with its character sequence reversed (not reverse complemented) called the 'mirror' index. To see how this helps, consider a matching policy that allows one mismatch in the alignment. A valid alignment with one mismatch falls into one of two cases according to which half of the read contains the mismatch. Bowtie proceeds in two phases corresponding to those two cases. Phase 1 loads the forward index into memory and invokes the aligner with the constraint that it may not substitute at positions in the query's right half. Phase 2 uses the mirror index and invokes the aligner on the reversed query, with the constraint that the aligner may not substitute at positions in the reversed query's right half (the original query's left half). The constraints on backtracking into the right half prevent excessive backtracking, whereas the use of two phases and two indices maintains full sensitivity.

Unfortunately, it is not possible to avoid excessive backtracking fully when alignments are permitted to have two or more mismatches. In our experiments, we have observed that excessive backtracking is significant only when a read has many low-quality positions and does not align or aligns poorly to the reference. These cases can trigger in excess of 200 backtracks per read because there are many legal combinations of low-quality positions to be explored before all possibilities are exhausted. We mitigate this cost by enforcing a limit on the number of backtracks allowed before a search is terminated (default: 125). The limit prevents some legitimate,

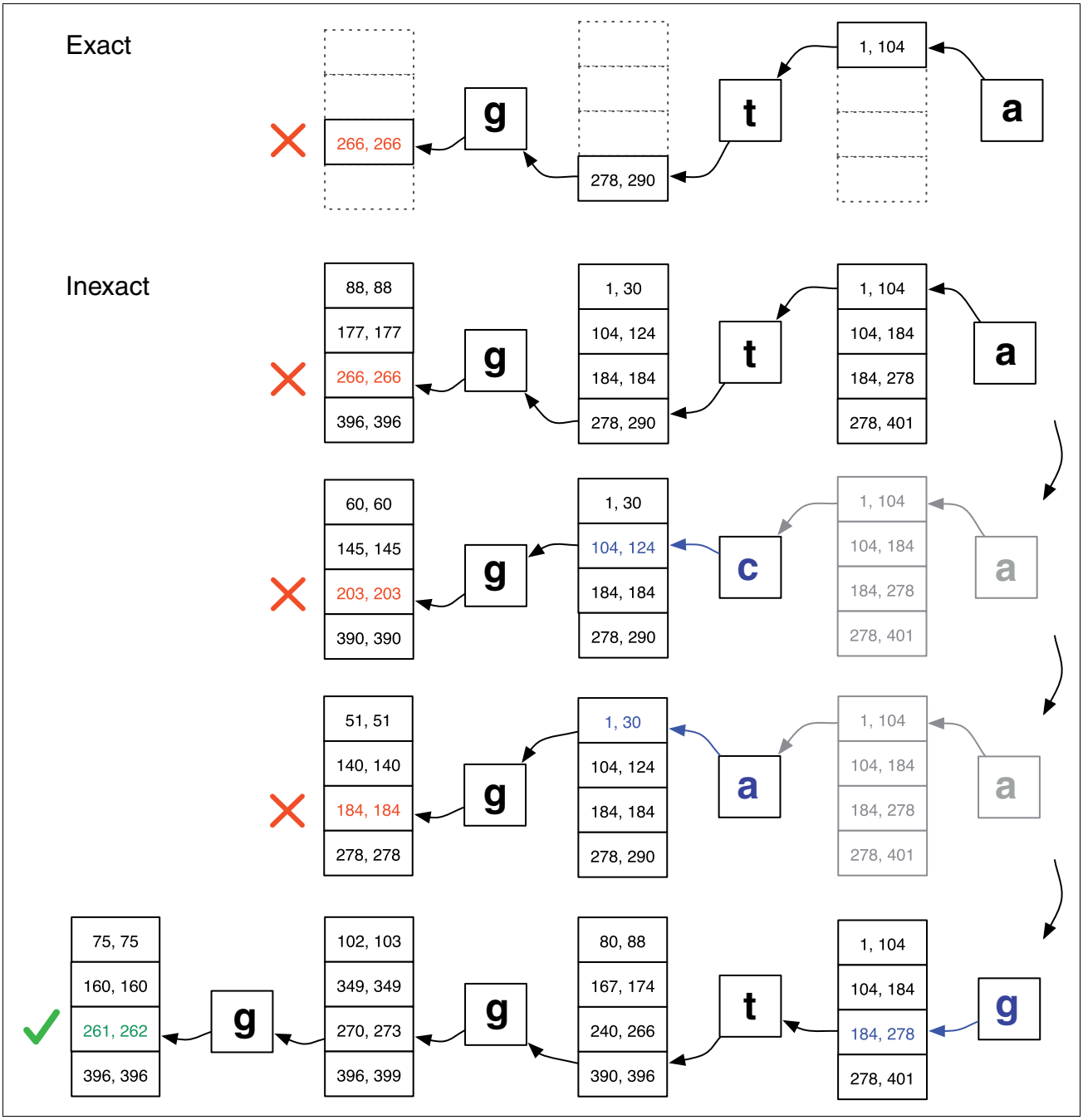


Figure 2
Exact matching versus inexact alignment. Illustration of how EXACTMATCH (top) and Bowtie's aligner (bottom) proceed when there is no exact match for query 'ggta' but there is a one-mismatch alignment when 'a' is replaced by 'g'. Boxed pairs of numbers denote ranges of matrix rows beginning with the suffix observed up to that point. A red X marks where the algorithm encounters an empty range and either aborts (as in EXACTMATCH) or backtracks (as in the inexact algorithm). A green check marks where the algorithm finds a nonempty range delimiting one or more occurrences of a reportable alignment for the query.

low-quality alignments from being reported, but we expect that this is a desirable trade-off for most applications.

Phased Maq-like search

Bowtie allows the user to select the number of mismatches permitted (default: two) in the high-quality end of a read (default: the first 28 bases) as well as the maximum acceptable quality distance of the overall alignment (default: 70). Quality values are assumed to follow the definition in PHRED [24], where p is the probability of error and Q = -10log p.

Both the read and its reverse complement are candidates for alignment to the reference. For clarity, this discussion considers only the forward orientation. See Additional data file 1 (Supplementary Discussion 2) for an explanation of how the reverse complement is incorporated.

The first 28 bases on the high-quality end of the read are termed the 'seed'. The seed consists of two halves: the 14 bp on the high-quality end (usually the 5' end) and the 14 bp on the low-quality end, termed the 'hi-half' and the 'lo-half', respectively. Assuming the default policy (two mismatches permitted in the seed), a reportable alignment will fall into one of four cases: no mismatches in seed (case 1); no mismatches in hi-half, one or two mismatches in lo-half (case 2); no mismatches in lo-half, one or two mismatches in hi-half (case 3); and one mismatch in hi-half, one mismatch in lo-half (case 4).

All cases allow any number of mismatches in the nonseed part of the read and all cases are also subject to the quality distance constraint.

The Bowtie algorithm consists of three phases that alternate between using the forward and mirror indices, as illustrated in Figure 3. Phase 1 uses the mirror index and invokes the aligner to find alignments for cases 1 and 2. Phases 2 and 3 cooperate to find alignments for case 3: Phase 2 finds partial alignments with mismatches only in the hi-half and phase 3 attempts to extend those partial alignments into full alignments. Finally, phase 3 invokes the aligner to find alignments for case 4.

Performance results

We evaluated the performance of Bowtie using reads from the 1,000 Genomes project pilot (National Center for Biotechnology Information [NCBI] Short Read Archive:SRR001115). A total of 8.84 million reads, about one lane of data from an Illumina instrument, were trimmed to 35 bp and aligned to the human reference genome [NCBI build 36.3]. Unless specified otherwise, read data are not filtered or modified (besides trimming) from how they appear in the archive. This leads to about 70% to 75% of reads aligning somewhere to the genome. In our experience, this is typical for raw data from the archive. More aggressive filtering leads to higher alignment rates and faster alignment.

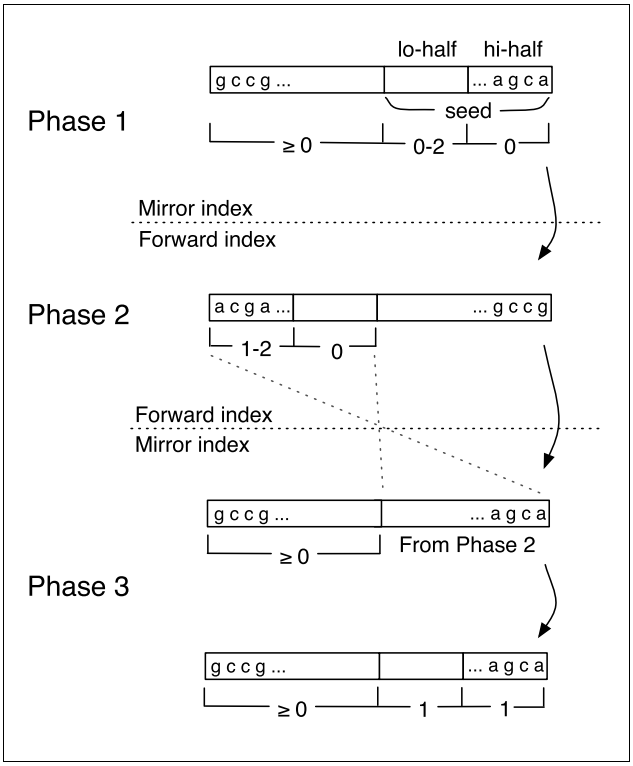


Figure 3
The three phases of the Bowtie algorithm for the Maq-like policy. A three-phase approach finds alignments for two-mismatch cases 1 to 4 while minimizing backtracking. Phase 1 uses the mirror index and invokes the aligner to find alignments for cases 1 and 2. Phases 2 and 3 cooperate to find alignments for case 3: Phase 2 finds partial alignments with mismatches only in the hi-half, and phase 3 attempts to extend those partial alignments into full alignments. Finally, phase 3 invokes the aligner to find alignments for case 4.

All runs were performed on a single CPU. Bowtie speedups were calculated as a ratio of wall-clock alignment times. Both wall-clock and CPU times are given to demonstrate that input/output load and CPU contention are not significant factors.

The time required to build the Bowtie index was not included in the Bowtie running times. Unlike competing tools, Bowtie can reuse a pre-computed index for the reference genome across many alignment runs. We anticipate most users will simply download such indices from a public repository. The Bowtie site [25] provides indices for current builds of the human, chimp, mouse, dog, rat, and *Arabidopsis thaliana* genomes, as well as many others.

Results were obtained on two hardware platforms: a desktop workstation with 2.4 GHz Intel Core 2 processor and 2 GB of RAM; and a large-memory server with a four-core 2.4 GHz AMD Opteron processor and 32 GB of RAM. These are denoted 'PC' and 'server', respectively. Both PC and server run Red Hat Enterprise Linux AS release 4.

Comparison to SOAP and Maq

Maq is a popular aligner [1,4,5,26,27] that is among the fastest competing open source tools for aligning millions of Illumina reads to the human genome. SOAP is another open source tool that has been reported and used in short-read projects [6,28].

Table 1 presents the performance and sensitivity of Bowtie v0.9.6, SOAP v1.10, and Maq v0.6.6. SOAP could not be run on the PC because SOAP's memory footprint exceeds the PC's physical memory. The 'soap.contig' version of the SOAP binary was used. For comparison with SOAP, Bowtie was invoked with '-v 2' to mimic SOAP's default matching policy (which allows up to two mismatches in the alignment and disregards quality values), and with '--maxns 5' to simulate SOAP's default policy of filtering out reads with five or more no-confidence bases. For the Maq comparison Bowtie is run with its default policy, which mimics Maq's default policy of allowing up to two mismatches in the first 28 bases and enforcing an overall limit of 70 on the sum of the quality values at all mismatched read positions. To make Bowtie's memory footprint more comparable to Maq's, Bowtie is invoked with the '-z' option in all experiments to ensure that only the forward or mirror index is resident in memory at one time.

The number of reads aligned indicates that SOAP (67.3%) and Bowtie -v 2 (67.4%) have comparable sensitivity. Of the reads aligned by either SOAP or Bowtie, 99.7% were aligned by both, 0.2% were aligned by Bowtie but not SOAP, and 0.1% were aligned by SOAP but not Bowtie. Maq (74.7%) and Bowtie (71.9%) also have roughly comparable sensitivity, although Bowtie lags by 2.8%. Of the reads aligned by either Maq or Bowtie, 96.0% were aligned by both, 0.1% were aligned by Bowtie but not Maq, and 3.9% were aligned by Maq but not Bowtie. Of the reads mapped by Maq but not Bowtie, almost all are due to a flexibility in Maq's alignment algorithm that allows some alignments to have three mismatches in the seed. The remainder of the reads mapped by Maq but not Bowtie are due to Bowtie's backtracking ceiling.

Maq's documentation mentions that reads containing 'poly-A artifacts' can impair Maq's performance. Table 2 presents performance and sensitivity of Bowtie and Maq when the read set is filtered using Maq's 'catfilter' command to eliminate poly-A artifacts. The filter eliminates 438,145 out of 8,839,010 reads. Other experimental parameters are identical to those of the experiments in Table 1, and the same observations about the relative sensitivity of Bowtie and Maq apply here.

Read length and performance

As sequencing technology improves, read lengths are growing beyond the 30-bp to 50-bp commonly seen in public databases today. Bowtie, Maq, and SOAP support reads of lengths up to 1,024, 63, and 60 bp, respectively, and Maq versions 0.7.0 and later support read lengths up to 127 bp. Table 3 shows performance results when the three tools are each used to align three sets of 2 M untrimmed reads, a 36-bp set, a 50-bp set and a 76-bp set, to the human genome on the server platform. Each set of 2 M is randomly sampled from a larger set (NCBI Short Read Archive: SRR003084 for 36-bp, SRR003092 for 50-bp, SRR003196 for 76-bp). Reads were sampled such that the three sets of 2 M have uniform per-base error rate, as calculated from per-base Phred qualities. All reads pass through Maq's 'catfilter'.

Bowtie is run both in its Maq-like default mode and in its SOAP-like '-v 2' mode. Bowtie is also given the '-z' option to ensure that only the forward or mirror index is resident in memory at one time. Maq v0.7.1 was used instead of Maq v0.6.6 for the 76-bp set because v0.6.6 cannot align reads longer than 63 bp. SOAP was not run on the 76-bp set because it does not support reads longer than 60 bp.

The results show that Maq's algorithm scales better overall to longer read lengths than Bowtie or SOAP. However, Bowtie in SOAP-like '-v 2' mode also scales very well. Bowtie in its default Maq-like mode scales well from 36-bp to 50-bp reads but is substantially slower for 76-bp reads, although it is still more than an order of magnitude faster than Maq.

Parallel performance

Alignment can be parallelized by distributing reads across concurrent search threads. Bowtie allows the user to specify a desired number of threads (option -p); Bowtie then launches the specified number of threads using the pthreads library. Bowtie threads synchronize with each other when fetching reads, outputting results, switching between indices, and performing various forms of global bookkeeping, such as marking a read as 'done'. Otherwise, threads are free to operate in parallel, substantially speeding up alignment on computers with multiple processor cores. The memory image of the index is shared by all threads, and so the footprint does not increase substantially when multiple threads are used. Table 4 shows performance results for running Bowtie v0.9.6 on the four-core server with one, two, and four threads.

Index building

Bowtie uses a flexible indexing algorithm [29] that can be configured to trade off between memory usage and running time. Table 5 illustrates this trade-off when indexing the entire human reference genome (NCBI build 36.3, contigs). Runs were performed on the server platform. The indexer was run four times with different upper limits on memory usage.

The reported times compare favorably with alignment times of competing tools that perform indexing during alignment. Less than 5 hours is required for Bowtie to both build and query a whole-human index with 8.84 million reads from the 1,000 Genome project (NCBI Short Read Archive:SRR001115) on a server, more than sixfold faster than the equivalent Maq run. The bottom-most row illus-

Table 3

Varying read length using Bowtie, Maq and SOAP

Length	Program	CPU time	Wall clock time	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
36 bp	Bowtie	6 m 15 s	6 m 21 s	1,305	-	62.2
	Maq	3 h 52 m 26 s	3 h 52 m 54 s	804	36.7×	65.0
	Bowtie -v 2	4 m 55 s	5 m 00 s	1,138	-	55.0
	SOAP	16 h 44 m 3 s	18 h 1 m 38 s	13,619	216×	55.1
50 bp	Bowtie	7 m 11 s	7 m 20 s	1,310	-	67.5
	Maq	2 h 39 m 56 s	2 h 40 m 9 s	804	21.8×	67.9
	Bowtie -v 2	5 m 32 s	5 m 46 s	1,138	-	56.2
	SOAP	48 h 42 m 4 s	66 h 26 m 53 s	13,619	691×	56.2
76 bp	Bowtie	18 m 58 s	19 m 6 s	1,323	-	44.5
	Maq 0.7.1	4 h 45 m 7 s	4 h 45 m 17 s	1,155	14.9×	44.9
	Bowtie -v 2	7 m 35 s	7 m 40 s	1,138	-	31.7

The performance of Bowtie v0.9.6, SOAP v1.10, and Maq versions v0.6.6 and v0.7.1 on the server platform when aligning 2 M untrimmed reads from the 1,000 Genome project (National Center for Biotechnology Information Short Read Archive: SRR003084 for 36 base pairs [bp], SRR003092 for 50 bp, and SRR003196 for 76 bp). For each read length, the 2 M reads were randomly sampled from the FASTQ file downloaded from the Archive such that the average per-base error rate as measured by quality values was uniform across the three sets. All reads pass through Maq's "catfilter". Maq v0.7.1 was used for the 76-bp reads because v0.6.6 does not support reads longer than 63 bp. SOAP is excluded from the 76-bp experiment because it does not support reads longer than 60 bp. Other experimental parameters are identical to those of the experiments in Table 1. CPU, central processing unit.

trates that the Bowtie indexer, with appropriate arguments, is memory-efficient enough to run on a typical workstation with 2 GB of RAM. Additional data file 1 (Supplementary discussions 3 and 4) explains the algorithm and the contents of the resulting index.

Software

Bowtie is written in C++ and uses the SeqAn library [30]. The converter to the Maq mapping format uses code from Maq.

Discussion

Bowtie exhibits a large performance advantage over both Maq and SOAP when mapping reads to the human genome. Bowtie's sensitivity in terms of reads aligned is comparable to that of SOAP and slightly less than Maq's, although the user may use command-line options to trade slower running time for greater sensitivity. Unlike SOAP, Bowtie's 1.3 GB memory footprint allows it to run on a typical PC with 2 GB of RAM. Bowtie aligns Illumina reads to the human genome at a rate of over 25 million reads per hour. Multiple processor cores can run parallel Bowtie threads to achieve even greater align-

Table 4

Bowtie parallel alignment performance

	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Speedup
Bowtie, one thread	18 m 19 s	18 m 46 s	28.3	1,353	-
Bowtie, two threads	20 m 34 s	10 m 35 s	50.1	1,363	1.77×
Bowtie, four threads	23 m 9 s	6 m 1 s	88.1	1,384	3.12×

Performance results for running Bowtie v0.9.6 on the four-core server with one, two, and four threads. Other experimental parameters are identical to those of the experiments in Table 1. CPU, central processing unit.

Table 5

Bowtie index building performance

Physical memory target (GB)	Actual peak memory footprint (GB)	Wall clock time
16	14.4	4 h 36 m
8	5.84	5 h 5 m
4	3.39	7 h 40 m
2	1.39	21 h 30 m

Performance results and memory footprints of running the Bowtie v0.9.6 indexer on the whole human genome (National Center for Biotechnology Information build 36.3, contigs). Runs were performed on the server platform. The indexer was run four times with different upper limits on memory usage. See Additional data file 1 (Supplementary Discussion 3 and Supplementary Table 1) for details.

ment speed; experiments show a speed up of 3.12 for four threads on a typical Opteron server.

Unlike many other short-read aligners, Bowtie creates a permanent index of the reference that may be re-used across alignment runs. Building the index is fast - Bowtie outperforms competing tools when aligning lanes of Illumina reads even with index construction time included. At 2.2 GB for the human genome, the on-disk size of a Bowtie index is small enough to distribute over the internet. The Bowtie website hosts pre-built indices for the human genome and several other model organisms including chimp, dog, rat, mouse, and chicken.

Bowtie's speed and small memory footprint are due chiefly to its use of the Burrows-Wheeler index in combination with the novel, quality-aware, backtracking algorithm introduced here. Double indexing is used to avoid the performance penalty of excessive backtracking.

Bowtie supports standard FASTQ and FASTA input formats, and comes with a conversion program that allows Bowtie output to be used with Maq's consensus generator and single nucleotide polymorphism caller.

Bowtie does not yet support paired-end alignment or alignments with insertions or deletions, although both improvements are planned for the future. Paired-end alignment is not difficult to implement in Bowtie's framework, and we expect that Bowtie's performance advantage will be comparable to, though perhaps somewhat less than, that of unpaired alignment mode. Support for insertions and deletions is also a conceptually straightforward addition.

Bowtie is free, open source software available from the Bowtie website [25].

Abbreviations

bp: base pair; BWT: Burrows-Wheeler transform; CPU: central processing unit; FM: full-text minute-space; GB: gigabytes; LF: last first; NCBI: National Center for Biotechnology Information.

Authors' contributions

BL developed the algorithms, collected results, and wrote most of the software. CT wrote some of the software. CT and MP contributed to discussions on algorithms. BL, CT, MP, and SLS wrote the manuscript.

Additional data files

The following additional data are included with the online version of this article: a document containing supplementary discussions, tables, and figures pertaining to algorithms for navigating the Burrows-Wheeler transform, the full four-phase version of the alignment algorithm that incorporates the reverse-complement, index construction, and components of the index (Additional data file 1).

Acknowledgements

The authors would like to thank Arthur Delcher for his thorough review of the manuscript, and Michael Schatz for helpful discussion of algorithms. This research was supported in part by NIH grants R01-LM006845 and R01-GM083873 to SLS.

References

1. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Backdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavare S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26:**779-785.
2. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316:**1497-1502.
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.
4. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.

5. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, *et al.*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456:**66-72.
6. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, *et al.*: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456:**60-65.
7. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.
8. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24:**713-714.
9. Kaiser J: **DNA sequencing. A plan to capture human diversity in 1000 genomes.** *Science* 2008, **319:**395.
10. Smith AD, Xuan Z, Zhang MQ: **Using quality scores and longer reads improves accuracy of Solexa read mapping.** *BMC Bioinformatics* 2008, **9:**128.
11. Lin H, Zhang Z, Zhang MQ, Ma B, Li M: **ZOOM! Zillions Of Oligos Mapped.** *Bioinformatics* 2008, **24:**2431-2437.
12. **SHRiMP - SHort Read Mapping Package** [http://compbio.cs.toronto.edu/shrimp/]
13. Baeza-Yates RA, Perleberg CH: **Fast and practical approximate string matching.** *Inf Process Lett* 1996, **59:**21-27.
14. Burkhardt S, Kärkkäinen J: **Better Filtering with Gapped q-Grams.** *Fundam Inf* 2003, **56:**51-70.
15. Ma B, Tromp J, Li M: **PatternHunter: faster and more sensitive homology search.** *Bioinformatics* 2002, **18:**440-445.
16. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
17. Burrows M, Wheeler DJ: *A Block Sorting Lossless Data Compression Algorithm. Technical Report 124* Palo Alto, CA: Digital Equipment Corporation; 1994.
18. Ferragina P, Manzini G: **Opportunistic data structures with applications.** [http://web.unipmn.it/~manzini/papers/focs00draft.pdf].
19. Ferragina P, Manzini G: **An experimental study of an opportunistic index.** In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete algorithms* Washington, DC: Society for Industrial and Applied Mathematics; 2001:269-278.
20. Healy J, Thomas EE, Schwartz JT, Wigler M: **Annotating large genomes with exact word matches.** *Genome Res* 2003, **13:**2306-2315.
21. Lippert RA: **Space-efficient whole genome comparisons with Burrows-Wheeler transforms.** *J Comput Biol* 2005, **12:**407-415.
22. Graf S, Nielsen FG, Kurtz S, Huynen MA, Birney E, Stunnenberg H, Flicek P: **Optimized design and assessment of whole genome tiling arrays.** *Bioinformatics* 2007, **23:**i195-i204.
23. Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM: **Compressed indexing and local alignment of DNA.** *Bioinformatics* 2008, **24:**791-797.
24. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8:**186-194.
25. **Bowtie: An ultrafast memory-efficient short read aligner** [http://bowtie.cbcb.umd.edu/]
26. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722-729.
27. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G: **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*.** *Nat Genet* 2008, **40:**987-993.
28. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320:**1344-1349.
29. Kärkkäinen J: **Fast BWT in small space by blockwise suffix sorting.** *Theor Comput Sci* 2007, **387:**249-257.
30. Doring A, Weese D, Rausch T, Reinert K: **SeqAn an efficient, generic C++ library for sequence analysis.** *BMC Bioinformatics* 2008, **9:**11.

A.6 Genomes

Recommended reading for Chapter 6 (listed in Section 6.11).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- International Human Genome Sequencing Consortium et al. (2001) — the human genome draft; what a genome looks like, and how it was obtained.
  DOI · via SU library (SU login)

  Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium et al. (2001). *Nature*. doi:10.1038/35057062

  > The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

- Lynch and Conery (2003) — why prokaryotic and eukaryotic genomes differ; the population-genetic theory of genome complexity.
  DOI · via SU library (SU login)

  The Origins of Genome Complexity. Lynch & Conery (2003). *Science*. doi:10.1126/science.1089370

  > Complete genomic sequences from diverse phylogenetic lineages reveal notable increases in genome complexity from prokaryotes to multicellular eukaryotes. The changes include gradual increases in gene number, resulting from the retention of duplicate genes, and more abrupt increases in the abundance of spliceosomal introns and mobile genetic elements. We argue that many of these modifications emerged passively in response to the long-term population-size reductions that accompanied increases in organism size. According to this model, much of the restructuring of eukaryotic genomes was initiated by non-adaptive processes, and this in turn provided novel substrates for the secondary evolution of phenotypic complexity by natural selection. The enormous long-term effective population sizes of prokaryotes may impose a substantial barrier to the evolution of complex genomes and morphologies.

- McInerney et al. (2017) — why prokaryotes have pan-genomes; core and accessory genes, and adaptive vs. neutral explanations.
  DOI · via SU library (SU login)

  Why prokaryotes have pangenomes. McInerney et al. (2017). *Nature Microbiology*. doi:10.1038/nmicrobiol.2017.40

- Green et al. (2010) — the draft Neanderthal genome and the discovery of admixture with modern humans.
  DOI · via SU library (SU login)

  A Draft Sequence of the Neandertal Genome. Green et al. (2010). *Science*. doi:10.1126/science.1188021

  > Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

- Bourque et al. (2018) — ten things you should know about transposable elements.
  full text below · DOI · open access (CC BY 4.0)

  Ten things you should know about transposable elements. Bourque et al. (2018). *Genome Biology*. doi:10.1186/s13059-018-1577-z

  > Transposable elements (TEs) are major components of eukaryotic genomes. However, the extent of their impact on genome evolution, function, and disease remain a matter of intense interrogation. The rise of genomics and large-scale functional assays has shed new light on the multi-faceted activities of TEs and implies that they should no longer be marginalized. Here, we introduce the fundamental properties of TEs and their complex interactions with their cellular environment, which are crucial to understanding their impact and manifold consequences for organismal biology. While we draw examples primarily from mammalian systems, the core concepts outlined here are relevant to a broad range of organisms.

- Eddy (2012) — the C-value paradox, junk DNA and ENCODE, in two pages.
  DOI · via SU library (SU login)

  The C-value paradox, junk DNA and ENCODE. Eddy (2012). *Current Biology*. doi:10.1016/j.cub.2012.10.002

The open-access papers are reproduced below under their original licences.

Bourque et al. Genome Biology (2018) 19:199
https://doi.org/10.1186/s13059-018-1577-z

Genome Biology

REVIEW Open Access

Ten things you should know about transposable elements

Guillaume Bourque[1,2*], Kathleen H. Burns[3], Mary Gehring[4], Vera Gorbunova[5], Andrei Seluanov[5], Molly Hammell[6], Michaël Imbeault[7], Zsuzsanna Izsvák[8], Henry L. Levin[9], Todd S. Macfarlan[9], Dixie L. Mager[10] and Cédric Feschotte[11*]

Abstract

Transposable elements (TEs) are major components of eukaryotic genomes. However, the extent of their impact on genome evolution, function, and disease remain a matter of intense interrogation. The rise of genomics and large-scale functional assays has shed new light on the multi-faceted activities of TEs and implies that they should no longer be marginalized. Here, we introduce the fundamental properties of TEs and their complex interactions with their cellular environment, which are crucial to understanding their impact and manifold consequences for organismal biology. While we draw examples primarily from mammalian systems, the core concepts outlined here are relevant to a broad range of organisms.

Transposable elements come in many different forms and shapes

Transposable elements (TEs) are DNA sequences that have the ability to change their position within a genome. As a result of their deep evolutionary origins and continuous diversification, TEs come in a bewildering variety of forms and shapes (Fig. 1). TEs can be divided into two major classes based on their mechanism of transposition, and each class can be subdivided into subclasses based on the mechanism of chromosomal integration. Class 1 elements, also known as retrotransposons, mobilize through a 'copy-and-paste' mechanism whereby a RNA intermediate is reverse-transcribed into a cDNA copy that is integrated elsewhere in the genome [1]. For long terminal repeat (LTR) retrotransposons, integration occurs by means of a cleavage and strand-transfer reaction catalyzed by an integrase much like retroviruses [2]. For non-LTR retrotransposons, which include both long and short interspersed nuclear elements (LINEs and SINEs), chromosomal integration is coupled to the reverse transcription through a process referred to as target-primed reverse transcription [3]. Class 2 elements, also known as DNA transposons, are mobilized via a DNA intermediate, either directly through a 'cut-and-paste' mechanism [4, 5] or, in the case of *Helitrons*, a 'peel-and-paste' replicative mechanism involving a circular DNA intermediate [6]. For detailed reviews on individual TE types and transposition mechanisms, we refer the reader to the monograph edited by Craig et al. [7].

Each TE subclass is further divided into subgroups (or superfamilies) that are typically found across a wide range of organisms, but share a common genetic organization and a monophyletic origin. For example, Ty3/*gypsy* and Ty1/*copia* elements are two major superfamilies of LTR retrotransposons that occur in virtually all major groups of eukaryotes [8]. Similarly, Tc1/*mariner*, hAT (hobo-Ac-Tam3), and MULEs (Mutator-like elements) are three superfamilies of DNA transposons that are widespread across the eukaryotic tree [9]. At the most detailed level of TE classification, elements are grouped into families or subfamilies, which can be defined as a closely related group of elements that can be traced as descendants of a single ancestral unit [10]. This ancestral copy can be inferred as a consensus sequence, which is representative of the entire (sub)family [11, 12]. Thus, in principle, every TE sequence in a genome can be affiliated to a (sub)family, superfamily, subclass, and class (Fig. 1). However, much like the taxonomy of species, the classification of TEs is in constant flux, perpetually subject to revision due to the discovery of completely novel TE types, the introduction of new levels of granularity in the classification, and ongoing development of methods and criteria to detect and classify TEs [13, 14].

* Correspondence: guil.bourque@mcgill.ca; cf458@cornell.edu
[1]Department of Human Genetics, McGill University, Montréal, Québec H3A 0G1, Canada
[11]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850, USA
Full list of author information is available at the end of the article

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

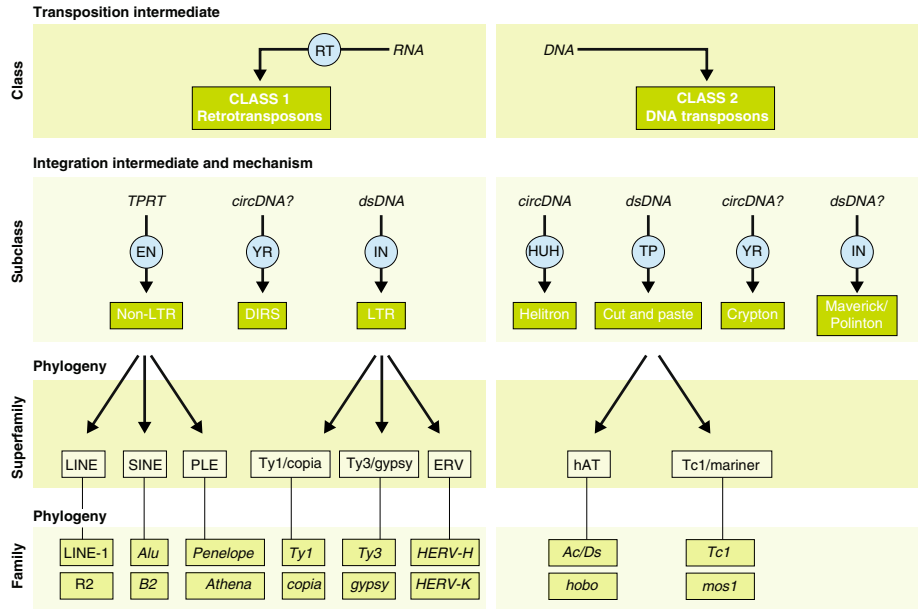


Fig. 1 Classification of eukaryotic transposable elements. Schematic and examples showing the key features and relationships between TE classes, subclasses, superfamilies, and families. Blue circles represent TE-encoded enzymes. circDNA circular DNA intermediate, DIRS Dictyostelium repetitive sequence, dsDNA linear double-stranded DNA intermediate, EN endonuclease, IN integrase, PLEs Penelope-like elements, HUH, Rep/Helicase protein with HUH endonuclease activity, RT reverse transcriptase, TP transposase, TPRT target primed reverse transcription, YR tyrosine recombinase (for other abbreviations, see text)

TEs are not randomly distributed in the genome

The genome may be viewed as an ecosystem inhabited by diverse communities of TEs, which seek to propagate and multiply through sophisticated interactions with each other and with other components of the cell [15]. These interactions encompass processes familiar to ecologists, such as parasitism, cooperation, and competition [16]. Thus, it is perhaps not surprising that TEs are rarely, if ever, randomly distributed in the genome. TEs exhibit various levels of preference for insertion within certain features or compartments of the genome (Fig. 2). These are often guided by opposite selective forces, a balancing act of facilitating future propagation while mitigating deleterious effects on host cell function. At the most extreme end of the site-selection spectrum, many elements have evolved mechanisms to target specific loci where their insertions are less detrimental to the host but favorable for their propagation [17]. For instance, several retrotransposons in species as diverse as slime mold and budding and fission yeast have evolved independently, but convergently, the capacity to target the upstream regions of genes transcribed by RNA polymerase III, where they do not appear to affect host gene expression but retain the ability to be transcribed themselves [17–20].

Natural selection and genetic drift are also powerful forces shaping the distribution and accumulation of TEs [21]. Insertions that are strongly deleterious are rapidly removed from the population. Insertions that have little or no effects on genome function and host fitness may reach fixation according to the efficiency of selection and drift at purging these insertions from the population, which vary greatly among species [21]. Selective forces can explain why some elements are more likely to be retained in certain genomic locations than others [22, 23]. For instance, de novo insertions of the human LINE 1 (L1) retrotransposon readily occur within (and disrupt) gene exons [24], but very few if any L1 elements have been fixed within the coding region of human genes [25]. Similarly, no LTR retrotransposon is known to exhibit insertion preference with regard to which DNA strand is transcribed, and yet these elements are strongly depleted in the sense orientation within human introns—most likely due to their propensity to interfere with gene splicing and polyadenylation when inserted in sense orientation [11, 26]. Perhaps because of some of these shared properties, the evolutionary trajectories of TE accumulation in mammals were found to be conserved across species in spite of clade specific differences in TE content. [27]. Thus, the success and diversity of TEs in a genome are shaped both by properties intrinsic to the elements as well as evolutionary forces acting at the level of the host species. A solid comprehension of how these forces act together is paramount to understanding the impact of TEs on organismal biology.

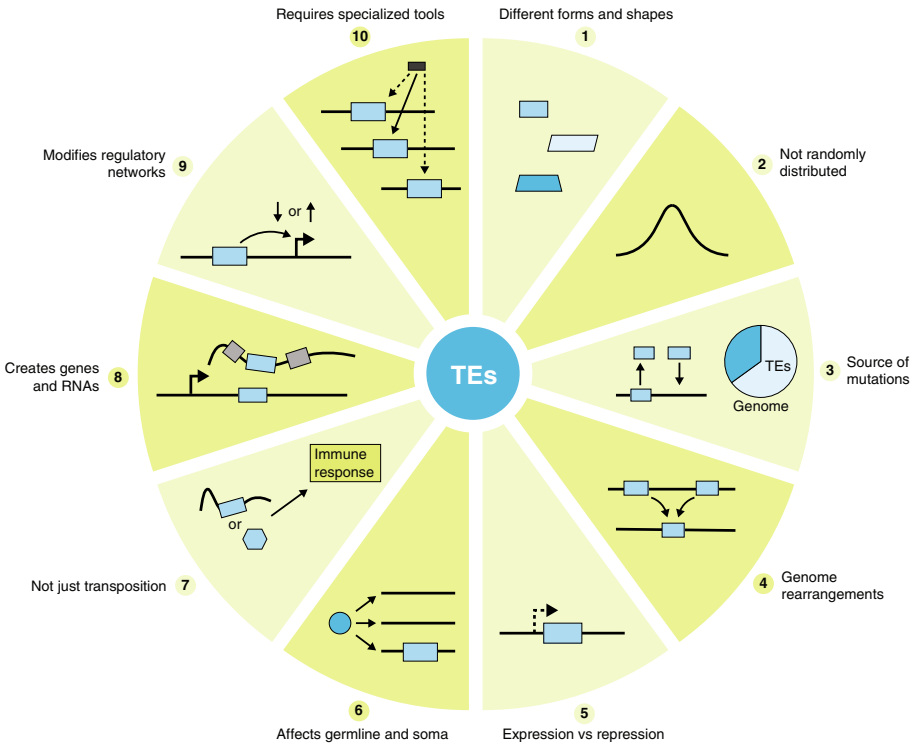


Fig. 2 Ten things you should know about transposable elements (TEs). Examples of how TEs can impact genomes in direct and indirect ways. Blue boxes represent TEs, gray boxes represent canonical exons, and the black box represents a sequencing read. Right-angled arrows represent gene or TE promoters

TEs are an extensive source of mutations and genetic polymorphisms

TEs occupy a substantial portion of the genome of a species, including a large fraction of the DNA unique to that species. In maize, where Barbara McClintock did her seminal work [28], an astonishing 60 to 70% of the genome is comprised of LTR retrotransposons, many of which are unique to this species or its close wild relatives, but the less prevalent DNA transposons are currently the most active and mutagenic [29–32] (Fig. 2). Similarly, the vast majority of TE insertions in *Drosophila melanogaster* are absent at the orthologous site in its closest relative *D. simulans* (and vice versa), and most are not fixed in the population [33, 34]. Many TE families are still actively transposing and the process is highly mutagenic; more than half of all known phenotypic mutants of *D. melanogaster* isolated in the laboratory are caused by spontaneous insertions of a wide variety of TEs [35]. Transposition events are also common and mutagenic in laboratory mice, where ongoing activity of several families of LTR elements are responsible for 10–15% of all inherited mutant phenotypes [36]. This contribution of TEs to genetic diversity may be underestimated, as TEs can be more active when organisms are under stress, such as in their natural environment [37, 38].

Because TE insertions rarely provide an immediate fitness advantage to their host, those reaching fixation in the population do so largely by genetic drift and are subsequently eroded by point mutations that accumulate neutrally [21]. Over time, these mutations result in TEs that can no longer encode transposition enzymes and produce new integration events. For instance, our (haploid) genome contains ~500,000 L1 copies, but more than 99.9% of these L1 copies are fixed and no longer mobile due to various forms of mutations and truncations [39, 40]. It is estimated that each person carries a set of ~100 active L1 elements, and most of these are young insertions still segregating within the human population [41–43]. Thus, as for any other organism, the 'reference' human genome sequence does not represent a comprehensive inventory of TEs in humans. Thousands of 'non-reference', unfixed TE insertions have been catalogued through whole genome sequencing and other targeted approaches [44]. On average, any two human haploid genomes differ by approximately a thousand TE insertions, primarily from the L1 or Alu families. The number of TE insertion polymorphisms in a species with much higher TE activity such as maize [32] dwarfs the number in humans.

If TEs bring no immediate benefit to their host and are largely decaying neutrally once inserted, how do they

persist in evolution? One key to this conundrum is the ability of TEs not only to propagate vertically but also horizontally between individuals and species. There is now a large body of evidence supporting the idea that horizontal transposon transfer is a common phenomenon that affects virtually every major type of TE and all branches of the tree of life [45, 46]. While the cellular mechanisms underlying horizontal transposon transfer remain murky, it is increasingly apparent that the intrinsic mobility of TEs and ecological interactions between their host species, including those with pathogens and parasites, facilitate the transmission of elements between widely diverged taxa [47–49].

TEs are associated with genome rearrangements and unique chromosome features

Transposition represents a potent mechanism of genome expansion that over time is counteracted by the removal of DNA via deletion. The balance between the two processes is a major driver in the evolution of genome size in eukaryotes [21, 50, 51]. Several studies have demonstrated the impact and range of this shuffling and cycling of genomic content on the evolution of plant and animal genomes [52–55]. Because the insertion and removal of TEs is often imprecise, these processes can indirectly affect surrounding host sequences. Some of these events occur at high enough frequency to result in vast amounts of duplication and reshuffling of host sequences, including genes and regulatory sequences. For example, a single group of DNA transposons (MULEs) has been responsible for the capture and reshuffling of ~ 1,000 gene fragments in the rice genome [56]. Such studies have led to the conclusion that the rate at which TEs transpose, which is in part under host control, is an important driver of genome evolution [57–59].

In addition to rearrangements induced as a byproduct of transposition, TEs can promote genomic structural variation long after they have lost the capacity to mobilize [60]. In particular, recombination events can occur between the highly homologous regions dispersed by related TEs at distant genomic positions and result in large-scale deletions, duplications, and inversions [59, 61–63] (Fig. 2). TEs also provide regions of microhomology that predispose to template switching during repair of replication errors leading to another source of structural variants [64]. These non-transposition-based mechanisms for TE-induced or TE-enabled structural variation have contributed substantially to genome evolution. These processes can also make the identification of actively transposing elements more difficult in population studies that infer the existence of active elements through the detection of non-reference insertions.

TEs also contribute to specialized chromosome features. An intriguing example is in *Drosophila*, where LINE-like retrotransposons form and maintain the telomeres in replacement of the telomerase enzyme which has been lost during dipteran evolution [65]. This domestication event could be viewed as a replay of what might have happened much earlier in eukaryotic evolution to solve the 'end problem' created by the linearization of chromosomes. Indeed, the reverse transcriptase component of telomerase is thought to have originated from an ancient lineage of retroelements [66, 67]. TE sequences and domesticated transposase genes also play structural roles at centromeres [68–70].

There is an intrinsic balance between TE expression and repression

To persist in evolution, TEs must strike a delicate balance between expression and repression (Fig. 2). Expression should be sufficient to promote amplification, but not so vigorous as to lead to a fitness disadvantage for the host that would offset the benefit to the TE of increased copy numbers. This balancing act may explain why TE-encoded enzymes are naturally suboptimal for transposition [71, 72] and why some TEs have evolved self-regulatory mechanisms controlling their own copy numbers [73, 74]. A variety of host factors are also employed to control TE expression, which includes a variety of small RNA, chromatin, and DNA modification pathways [75–78], as well as sequence-specific repressors such as the recently profiled KRAB zinc-finger proteins [79–82]. However, many of these silencing mechanisms must be at least partially released to permit developmental regulation of host gene expression programs, particularly during early embryonic development. For example, genome-wide loss of DNA methylation is necessary to reset imprinted genes in primordial germ cells [83]. This affords TEs an opportunity, as reduced DNA methylation often promotes TE expression. Robust expression of a TE in the germ lineage (but not necessarily in the gametes themselves) is often its own downfall. In one example of a clever trick employed by the host, TE repression is relieved in a companion cell derived from the same meiotic product as flowering plant sperm [84]. However, this companion cell does not contribute genetic material to the next generation. Thus, although TEs transpose in a meiotic product, the events are not inherited. Instead, TE activity in the companion cell may further dampen TE activity in sperm via the import of TE-derived small RNAs [85].

Another important consequence of the intrinsic expression/repression balance is that the effects of TEs on a host can vary considerably among tissue types and stages of an organism's life cycle. From the TE's perspective, an ideal scenario is to be expressed and active in the germline, but not in the soma, where expression would gain the TE no advantage, only disadvantage [86].

This is indeed observed among many species, with ciliates representing an extreme example of this division—TEs are actively deleted from the somatic macronucleus but retained in the micronucleus, or germline [87]. Another example is the P-elements in *Drosophila*, which are differentially spliced in the germline versus soma [88]. Many organisms, including plants, do not differentiate germ lineage cells early in development; rather, they are specified from somatic cells shortly before meiosis commences. Thus, TEs that transpose in somatic cells in plants have the potential to be inherited, which suggests that the interest of TEs and host are in conflict across many more cells and tissues than in animals with a segregated germline.

TEs are insertional mutagens in both germline and soma

Like other species, humans contend with a contingent of currently active TEs where the intrinsic balance between expression and repression is still at play [89]. For us, this includes L1 and other mobile elements that depend on L1-encoded proteins for retrotransposition [90, 91]. These elements are responsible for new germline insertions that can cause genetic disease. More than 120 independent TE insertions have been associated with human disease [24]. The rate of de novo germline transposition in humans is approximately one in 21 births for *Alu* [92] and one in 95 births for L1 [93].

Historically, little attention has been given to transposition in somatic cells and its consequences, because somatic transposition may be viewed as an evolutionary dead-end for the TE with no long-term consequences for the host species. Yet, there is abundant evidence that TEs are active in somatic cells in many organisms [94] (Fig. 2). In humans, L1 expression and transposition have been detected in a variety of somatic contexts, including early embryos and certain stem cells [95, 96]. There is also a great deal of interest in mobile element expression and activity in the mammalian brain, where L1 transposition has been proposed to diversify neuronal cell populations [97–99]. One challenge for assessing somatic activity has rested with the development of reliable single cell insertion site mapping strategies [100–103].

Somatic activity has also been observed in human cancers, where tumors can acquire hundreds of new L1 insertions [104–109]. Just like for human polymorphisms, somatic activity in human cancers is caused by small numbers of so-called 'hot' L1 loci [41, 107]. The activities of these master copies varies depending on the individual [105], tumor type [105], and timeframe in the clonal evolution of the tumor [106, 110]. Some of these de novo L1 insertions disrupt critical tumor suppressors and oncogenes and thus drive cancer formation [107], although the vast majority appear to be 'passenger' mutations [111]. Host cells have evolved several mechanisms to keep TEs in check. However, as the force of natural selection begins to diminish with age and completely drops in post-reproductive life, TEs may become more active [112].

TEs can be damaging in ways that do not involve transposition

TEs are best known for their mobility, in other words their ability to transpose to new locations. While the breakage and insertion of DNA associated with transposition represents an obvious source of cell damage, this is not the only or perhaps even the most common mechanism by which TEs can be harmful to their host. Reactivated transposons harm the host in multiple ways. First, de-repression of transposon loci, including their own transcription, may interfere with transcription or processing of host mRNAs through a myriad of mechanisms [113–115]. Genome-wide transcriptional de-repression of TEs has been documented during replicative senescence of human cells [116] and several mouse tissues, including liver, muscle, and brain [117, 118]. De-repression of LTR and L1 promoters can also cause oncogene activation in cancer [119]. Second, TE-encoded proteins such as the endonuclease activity of L1 ORF2p can induce DNA breaks and genomic instability [120]. Third, accumulation of RNA transcripts and extrachromosomal DNA copies derived from TEs may trigger an innate immune response leading to autoimmune diseases and sterile inflammation (Fig. 2). Activation of interferon response is now a well-documented property of transcripts derived from endogenous retroviruses and may give immunotherapies a boost in identifying and attacking cancer cells [121–123]. The relative contribution of all the above mechanisms in organismal pathologies remains to be determined.

Following transcription (and sometimes splicing) of TEs, the next step in the process involves translation of the encoded proteins and, for retroelements, reverse transcription of the TEs into cDNA substrates suitable for transposition. Once engaged by a TE-encoded reverse transcriptase protein, the resulting cytosolic DNAs and RNA:DNA hybrids can alert inflammatory pathways. An example of this is seen in patients with Aicardi–Goutières syndrome, where accumulation of TE-derived cytosolic DNA is due to mutations in pathways that normally block TE processing or degrade TE-derived DNA [124, 125]. Although not all TEs encode functional proteins, some do, including a few endogenous retroviruses capable of producing Gag, Pol, or envelope (Env) proteins [126]. Overexpression of these Env proteins can be cytotoxic, and has been linked to at least two neurodegenerative diseases, multiple sclerosis [127] and amytrophic lateral sclerosis [128]. Small accessory proteins produced by the youngest human endogenous retrovirus

(HERV) group, HERV-K (HML-2), may play a role in some cancers but the evidence remains circumstantial [129, 130].

A number of key coding and non-coding RNAs are derived from TEs

Although usually detrimental, there is growing evidence that TE insertions can provide raw material for the emergence of protein-coding genes and non-coding RNAs, which can take on important and, in some cases essential, cellular function [131–133] (Fig. 2). The process of TE gene 'domestication' or exaptation over evolutionary time contributes to both deeply conserved functions and more recent, species-specific traits. Most often, the ancestral or a somewhat modified role of a TE-encoded gene is harnessed by the host and conserved, while the rest of the TE sequence, and hence its ability to autonomously transpose, has been lost. Spectacular examples of deeply conserved TE-derived genes are *Rag1* and *Rag2*, that catalyze V(D)J somatic recombination in the vertebrate immune system. Both genes, and probably the DNA signals they recognize, were derived from an ancestral DNA transposon around 500 million years ago [134, 135]. Indeed, DNA transposases have been co-opted multiple times to form new cellular genes [70, 113].

The *gag* and *env* genes of LTR retrotransposons or endogenous retroviruses (ERVs) have also been domesticated numerous times to perform functions in placental development, contribute to host defense against exogenous retroviruses, act in brain development, and play other diverse roles [132, 136]. One of the most intriguing examples of TE domestication is the repeated, independent capture of ERV *env* genes, termed *syncytins*, which appear to function in placentation by facilitating cell–cell fusion and syncytiotrophoblast formation [137–139]. Notably, one or more such *syncytin* genes have been found in virtually every placental mammalian lineage where they have been sought, strongly suggesting that ERVs have played essential roles in the evolution and extreme phenotypic variability of the mammalian placenta. Another example of a viral-like activity re-purposed for host cell function is provided by the neuronal *Arc* gene, which arose from the *gag* gene from a LTR retrotransposon domesticated in the common ancestor of tetrapod vertebrates [140]. Genetic and biochemical studies of murine Arc show that it is involved in memory and synaptic plasticity and has preserved most of the ancestral activities of Gag, including the packaging and intercellular trafficking of its own RNA [140]. Remarkably, flies appear to have independently evolved a similar system of trans-synaptic RNA delivery involving a *gag*-like protein derived from a similar yet distinct lineage of LTR retrotransposons [141]. Thus, the biochemical activities of TE-derived proteins have been repeatedly co-opted during evolution to foster the emergence of convergent cellular innovations in different organisms.

TEs can donate their own genes to the host, but they can also add exons and rearrange and duplicate existing host genes. In humans, intronic *Alu* elements are particularly prone to be captured as alternative exons through cryptic splice sites residing within their sequences [142, 143]. L1 and SVA (SINE/VNTR/Alu) elements also contribute to exon shuffling through transduction events of adjacent host sequences during their mobilization [144, 145]. The reverse transcriptase activity of retroelements is also responsible for the trans-duplication of cellular mRNAs to create 'processed' retrogenes in a wide range of organisms [146, 147]. The L1 enzymatic machinery is thought to be involved in the generation of tens of thousands of retrogene copies in mammalian genomes, many of which remain transcribed and some of which have acquired new cellular functions [147, 148]. This is a process still actively shaping our genomes; it has been estimated that 1 in every 6000 humans carries a novel retrogene insertion [93].

TEs also make substantial contributions to non-protein coding functions of the cell. They are major components of thousands of long non-coding RNAs in human and mouse genomes, often transcriptionally driven by retroviral LTRs [149]. Some of these TE-driven lncRNAs appear to play important roles in the maintenance of stem cell pluripotency and other developmental processes [150–154]. Many studies have demonstrated that TE sequences embedded within lncRNAs and mRNAs can directly modulate RNA stability, processing, or localization with important regulatory consequences [114, 155–158]. Furthermore, TE-derived microRNAs [159] and other small RNAs processed from TEs [160] can also adopt regulatory roles serving host cell functions. The myriad of mechanisms by which TEs contribute to coding and non-coding RNAs illustrate the multi-faceted interactions between these elements and their host.

TEs contribute cis-regulatory DNA elements and modify transcriptional networks

Cis-regulatory networks coordinate the transcription of multiple genes that function in concert to orchestrate entire pathways and complex biological processes. In line with Barbara McClintock's insightful predictions [28], there is now mounting evidence that TEs have been a rich source of material for the modulation of eukaryotic gene expression (Fig. 2). Indeed, TEs can disperse vast amounts of promoters and enhancers [161–166], transcription factor binding sites [167–172], insulator sequences [173–175], and repressive elements

[176, 177] (reviewed in [178]). The varying coat colors of agouti mice provides a striking example of a host gene controlling coat color whose expression can be altered by the methylation levels of a TE upstream of its promoter [179, 180]. In the oil palm, the methylation level of a TE that sits within a gene important for flowering ultimately controls whether or not the plants bear oil-rich fruit [181].

As TE families typically populate a genome as a multitude of related copies, it has long been postulated that they have the potential to donate the same cis-regulatory module to 'wire' batteries of genes dispersed throughout the genome [182]. An increasing number of studies support this model and suggest that TEs have provided the building blocks for the assembly and remodeling of cis-regulatory networks during evolution, including pathways underlying processes as diverse as pregnancy [183, 184], stem cell pluripotency [150, 151, 171], neocortex development [185], innate immunity in mammals [163], or the response to abiotic stress in maize [186]. Indeed, TE sequences harbor all the necessary features of a 'classical' gene regulatory network [113, 114]. They are bound by diverse sets of transcription factors [172] integrate multiple inputs (activation/repression), respond to signals in both *cis* and *trans*, and are capable of co-ordinately regulating gene expression. In this context, TEs are highly suitable agents to modify biological processes by creating novel cis-regulatory circuits and fine-tuning pre-existing networks.

Analyzing TEs requires specialized tools

TEs have been historically neglected and remain frequently ignored in genomic studies in part because of their repetitive nature, which poses a number of analytical challenges and often requires the use of specialized tools [187]. As genomes can harbor thousands of copies of very similar TE sequences, uniqueness or, alternatively, repetitiveness of substrings within these regions need to be taken into consideration during both experimental design and analysis. As an example, short DNA oligos targeting a specific TE instance in the genome for PCR, short hairpin RNA, or CRISPR-Cas9 have to be carefully designed and validated to ensure that they are truly specific and target unique regions of the genome. In some scenarios, it can be acceptable or even desirable to target many elements simultaneously [150] or an entire TE family [153, 188–191].

Similarly, uniqueness and repetitiveness are important concepts to consider when aligning reads from next generation sequencing and analyzing TEs (Fig. 2). Various strategies exist to assign reads that could originate from multiple genomic locations: 1) mapping reads to consensus sequences of TE subfamilies [172]; 2) mapping to the genome and keeping only uniquely-mapping reads [163, 168]; 3) assigning multiple mapping reads at random between possible candidates [192]; or 4) redistributing them according to various algorithms, such as maximum likelihood [193, 194]. The choice is ultimately guided by the technique (such as ChIP-seq and RNA-seq) and the purpose of the analysis—is information about individual TE instances needed, or is a high-level tally of results for each subfamily sufficient? Notably, these issues of uniqueness will differ substantially depending on the species studied and the presence or absence of recently, or currently, active TE families. For example, mapping reads to TEs in the human genome will be less challenging than in the mouse genome given the more recent and mobile TE landscape of the latter species [36]. Finally, as sequencing technology and bioinformatics pipelines improve, notably with the increasing length of sequencing reads, many of the hurdles faced by earlier studies will be progressively removed [187].

Outlook

As potent insertional mutagens, TEs can have both positive and negative effects on host fitness, but it is likely that the majority of TE copies in any given species—and especially those such as humans with small effective population size—have reached fixation through genetic drift alone and are now largely neutral to their host. When can we say that TEs have been co-opted for cellular function? The publication of the initial ENCODE paper [195], which asserted 'function for 80% of the genome', was the subject of much debate and controversy. Technically speaking, ENCODE assigned only 'biochemical' activity to this large fraction of the genome. Yet critics objected to the grand proclamations in the popular press (The Washington Post Headline: "Junk DNA concept debunked by new analysis of the human genome") and to the ENCODE consortium's failure to prevent this misinterpretation [196–198]. To these critics, ignoring evolutionary definitions of function was a major misstep.

This debate can be easily extended to include TEs. TEs make up the vast majority of what is often referred to as 'junk DNA'. Today, the term is mostly used (and abused) by the media, but it has in fact deep roots in evolutionary biology [199]. Regardless of the semantics, what evidence is needed to assign a TE with a function? Many TEs encode a wide range of biochemical activities that normally benefit their own propagation. For example, TEs often contain promoter or enhancer elements that highjack cellular RNA polymerases for transcription and autonomous elements encode proteins with various biochemical and enzymatic activities, all of which are necessary for the transposon to replicate. Do these activities make them functional?

The vast differences in TEs between species make standard approaches to establish their regulatory roles particularly challenging [200]. For example, intriguing studies on the impact of HERVs, in particular HERV-H, in stem cells and pluripotency [150–152] must be interpreted using novel paradigms that do not invoke deep evolutionary conservation to imply function, as these particular ERVs are absent outside of great apes. Evolutionary constraint can be measured at shorter time scales, including the population level, but this remains a statistically challenging task especially for non-coding sequences. Natural loss-of-function alleles may exist in the human population and their effect on fitness can be studied if their impact is apparent, but these are quite rare and do not allow systematic studies. It is possible to engineer genetic knockouts of a particular human TE locus to test its regulatory role but those are restricted to in-vitro systems, especially when the orthologous TE does not exist in the model species. In this context, studying the impact of TEs in model species with powerful genome engineering tools and vast collections of mutants and other genetic resources, such as plants, fungi, and insects, will also continue to be extremely valuable.

Finally, a growing consensus is urging more rigor when assigning cellular function to TEs, particularly for the fitness benefit of the host [178]. Indeed, a TE displaying biochemical activity (such as those bound by transcription factors or lying within open chromatin regions) cannot be equated to a TE that shows evidence of purifying selection at the sequence level or, when genetically altered, result in a deleterious or dysfunctional phenotype. Recent advances in editing and manipulating the genome and the epigenome en masse yet with precision, including repetitive elements [153, 154, 189–191], offer the promise for a systematic assessment of the functional significance of TEs.

Abbreviations

Env: Envelope protein; ERV: Endogenous retrovirus; HERV: Human endogenous retrovirus; L1: Long interspersed nuclear element 1; LINE: Long interspersed nuclear element; LTR: Long terminal repeat; SINE: Short interspersed nuclear element; TE: Transposable element

Acknowledgments

We apologize to our colleagues whose relevant work and original articles could not be cited owing to space constraint. We thank Patricia Goerner-Potvin and Edward Chuong for early discussions about the manuscript. We are also indebted to Ayana Moosa and Susan Mahon from the McGill Bellairs Research Institute for their help in organizing the workshop that led to the concept of this review.

Funding

This work was supported by grants from the Canadian Institute for Health Research (CIHR-MOP-115090 and CIHR-CEE-151618). GB is supported by the Fonds de Recherche Santé Québec (FRSQ-25348 and FRSQ-35279). ZI is supported by ERC-2011-AdG 294742. CF, HLL, KHB, VG, and MG acknowledge funding by the National Institutes of Health. MG is also funded by the National Science Foundation.

Authors' contributions

All authors participated in the planning and the writing of this review. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Human Genetics, McGill University, Montréal, Québec H3A 0G1, Canada. [2]Canadian Center for Computational Genomics, McGill University, Montréal, Québec H3A 0G1, Canada. [3]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [4]Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [5]Department of Biology, University of Rochester, Rochester, NY 14627, USA. [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [7]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [8]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 13125 Berlin, Germany. [9]The Eunice Kennedy Shriver National Institute of Child Health and Human Development, The National Institutes of Health, Bethesda, Maryland, USA. [10]Terry Fox Laboratory, British Columbia Cancer Agency and Department of Medical Genetics, University of BC, Vancouver, BC V5Z1L3, Canada. [11]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850, USA.

Published online: 19 November 2018

References

1. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. Cell. 1985;40:491–500.
2. Brown PO, Bowerman B, Varmus HE, Bishop JM. Correct integration of retroviral DNA in vitro. Cell. 1987;49:347–56.
3. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72:595–605.
4. Greenblatt IM, Brink RA. Transpositions of modulator in maize into divided and undivided chromosome segments. Nature. 1963;197:412–3.
5. Rubin GM, Kidwell MG, Bingham PM. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell. 1982;29:987–94.
6. Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun. 2016;7:10716.
7. Craig NL, Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB. Mobile DNA III. 3rd ed. Washington, DC: American Society for Microbiology (ASM); 2015.
8. Malik HS, Eickbush TH. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. 2001;11:1187–97.
9. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–68.
10. Britten RJ, Kohne DE. Repeated sequences in DNA. Science. 1968;161:529–40.
11. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 1999;9:657–63.
12. Jurka J, Smith T. A fundamental division in the Alu family of repeated sequences. Proc Natl Acad Sci U S A. 1988;85:4775–8.
13. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
14. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017;8:19.
15. Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. 2009;25:317–23.

16. Robillard É, Rouzic AL, Zhang Z, Capy P, Hua-Van A. Experimental evolution reveals hyperparasitic interactions among transposable elements. Proc Natl Acad Sci U S A. 2016;113:14763–8.
17. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017; 18:292–308.
18. Spaller T, Kling E, Glöckner G, Hillmann F, Winckler T. Convergent evolution of tRNA gene targeting preferences in compact genomes. Mob DNA. 2016; 7:17.
19. Cheung S, Manhas S, Measday V. Retrotransposon targeting to RNA polymerase III-transcribed genes. Mob DNA. 2018;9:14.
20. Guo Y, Singh PK, Levin HL. A long terminal repeat retrotransposon of *Schizosaccharomyces japonicus* integrates upstream of RNA pol III transcribed genes. Mob DNA. 2015;6:19.
21. Lynch M. The origins of genome architecture. 1st ed. Sunderland: Sinauer Associates; 2007.
22. Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. PLOS Comput Biol. 2016;12:e1004956.
23. Kvikstad EM, Makova KD. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. Genome Res. 2010; 20:600–13.
24. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7:9.
25. Gotea V, Makalowski W. Do transposable elements really contribute to proteomes? Trends Genet. 2006;22:260–7.
26. Medstrand P, Van De Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 2002;12:1483–95.
27. Buckley RM, Kortschak RD, Raison JM, Adelson DL. Similar evolutionary trajectories for retrotransposon accumulation in mammals. Genome Biol Evol. 2017;9:2336–53.
28. McClintock B. Controlling elements and the gene: Cold Spring Harb Symp Quant Biol. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1956. p. 197–216.
29. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
30. Lazarow K, Doll M-L, Kunze R. Molecular Biology of Maize Ac/Ds elements: an overview. In: Peterson T, editor. Plant transposable elements. Totowa: Humana Press; 2013. p. 59–82.
31. Liu S, Yeh C-T, Ji T, Ying K, Wu H, Tang HM, et al. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 2009;5: e1000733.
32. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet. 2018;50:1282–8.
33. Kofler R, Nolte V, Schlötterer C. Tempo and mode of transposable element activity in Drosophila. PLOS Genet. 2015;11:e1005406.
34. Rahman R, Chirn G, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, et al. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. Nucleic Acids Res. 2015;43:10655–72.
35. Eickbush TH, Furano AV. Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev. 2002;12:669–74.
36. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, Van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2006;2:e2.
37. Lanciano S, Mirouze M. Transposable elements: all mobile, all different, some stress responsive, some adaptive? Curr Opin Genet Dev. 2018;49:106–14.
38. Horváth V, Merenciano M, González J. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet. 2017;33:832–41.
39. Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. Annu Rev Genet. 2001;35:501–38.
40. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol. 2002;3:research0052–1.
41. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100:5280–5.
42. Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, et al. Many human L1 elements are capable of retrotransposition. Nat Genet. 1997;16:37–43.
43. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, et al. LINE-1 retrotransposition activity in human genomes. Cell. 2010;141:1159–70.
44. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
45. Gilbert C, Feschotte C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev. 2018;49:15–24.
46. Wallau GL, Vieira C, Loreto ÉLS. Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. Mob DNA. 2018;9:6.
47. Gilbert C, Cordaux R. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. Curr Opin Virol. 2017;25:16–22.
48. Metzger MJ, Paynter AN, Siddall ME, Goff SP. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc Natl Acad Sci U S A. 2018;115:E4227–35.
49. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 2018;19:85.
50. Petrov DA. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 2002;61:531–44.
51. Schubert I, Vu GTH. Genome stability and evolution: attempting a holistic view. Trends Plant Sci. 2016;21:749–57.
52. Gregory TR, Johnston JS. Genome size diversity in the family Drosophilidae. Heredity. 2008;101:228–38.
53. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017;114:E1460–9.
54. Pellicer J, Kelly LJ, Leitch IJ, Zomlefer WB, Fay MF. A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family Melanthiaceae. New Phytol. 2014;201:1484–97.
55. Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, et al. Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. Genome Res. 2018;28:448–59.
56. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
57. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10:691–703.
58. Freeling M, Xu J, Woodhouse M, Lisch D. A solution to the C-value paradox and the function of junk DNA: the genome balance hypothesis. Mol Plant. 2015;8:899–910.
59. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65:505–30.
60. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38.
61. Deininger PL, Moran JV, Batzer MA, Kazazian HH. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 2003;13:651–8.
62. Ade C, Roy-Engel AM, Deininger PL. Alu elements: an intrinsic source of human genome instability. Curr Opin Virol. 2013;3:639–45.
63. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A. 2008;105:19366–71.
64. Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007;131:1235–47.
65. Pardue M-L, DeBaryshe PG. Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci U S A. 2011;108:20317–24.
66. Belfort M, Curcio MJ, Lue NF. Telomerase and retrotransposons: reverse transcriptases that shaped genomes. Proc Natl Acad Sci U S A. 2011;108: 20304–10.
67. Fulcher N, Derboven E, Valuchova S, Riha K. If the cap fits, wear it: an overview of telomeric structures over evolution. Cell Mol Life Sci. 2014; 71:847–65.
68. Casola C, Hucks D, Feschotte C. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol. 2007;25:29–41.
69. Kursel LE, Malik HS. Centromeres. Curr Biol. 2016;26:R487–90.
70. Jangam D, Feschotte C, Betrán E. Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet. 2017;33:917–31.

71. Lampe DJ, Akerley BJ, Rubin EJ, Mekalanos JJ, Robertson HM. Hyperactive transposase mutants of the Himar1 mariner transposon. Proc Natl Acad Sci U S A. 1999;96:11428–33.
72. Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. Nat Genet. 2009;41:753–61.
73. Lohe AR, Hartl DL. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. Mol Biol Evol. 1996;13:549–55.
74. Saha A, Mitchell JA, Nishida Y, Hildreth JE, Ariberre JA, Gilbert WV, et al. A trans-dominant form of gag restricts Ty1 retrotransposition and mediates copy number control. J Virol. 2015;89:3922–38.
75. Molaro A, Malik HS. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Curr Opin Genet Dev. 2016;37:51–8.
76. Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2017;553:228–32.
77. Goodier JL. Restricting retrotransposons: a review. Mob DNA. 2016;7:16.
78. Berrens RV, Andrews S, Spensberger D, Santos F, Dean W, Gould P, et al. An endosiRNA-based repression mechanism counteracts transposon activation during global dna demethylation in embryonic stem cells. Cell Stem Cell. 2017;21:694–703.e7.
79. Imbeault M, Trono D. As time goes by: KRABs evolve to KAP endogenous retroelements. Dev Cell. 2014;31:257–8.
80. Imbeault M, Helleboid P-Y, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature. 2017;543:550–4.
81. Yang P, Wang Y, Macfarlan TS. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. Trends Genet. 2017;33:871–81.
82. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. Development. 2017;144:2719–29.
83. Miyoshi N, Stel JM, Shioda K, Qu N, Odahima J, Mitsunaga S, et al. Erasure of DNA methylation, genomic imprints, and epimutations in a primordial germ-cell model derived from mouse pluripotent stem cells. Proc Natl Acad Sci U S A. 2016;113:9545–50.
84. Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136:461–72.
85. Martínez G, Panda K, Köhler C, Slotkin RK. Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. Nat Plants. 2016;2:16030.
86. Haig D. Transposable elements: self-seekers of the germline, team-players of the soma. BioEssays. 2016;38:1158–66.
87. Allen SE, Nowacki M. Necessity is the mother of invention: ciliates, transposons, and transgenerational inheritance. Trends Genet. 2017;33:197–207.
88. Teixeira FK, Okuniewska M, Malone CD, Coux R-X, Rio DC, Lehmann R. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. Nature. 2017;552:268–72.
89. Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet. 2012;46:651–75.
90. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet. 2011;12:187–215.
91. Burns KH, Boeke JD. Human transposon tectonics. Cell. 2012;149:740–52.
92. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, et al. Mobile elements create structural variation analysis of a complete human genome. Genome Res. 2009;19:1516–26.
93. Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010;20:1262–70.
94. Kazazian HH. Mobile DNA transposition in somatic cells. BMC Biol. 2011;9:62.
95. Garcia-Perez JL, Marchetto MC, Muotri AR, Coufal NG, Gage FH, O'shea KS, et al. LINE-1 retrotransposition in human embryonic stem cells. Hum Mol Genet. 2007;16:1569–77.
96. Klawitter S, Fuchs NV, Upton KR, Munoz-Lopez M, Shukla R, Wang J, et al. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat Commun. 2016;7:10286.
97. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.
98. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci. 2014;15:497–506.
99. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435:903–10.
100. Evrony GD, Lee E, Park PJ, Walsh CA. Resolving rates of mutation in the brain using single-neuron genomics. Elife. 2016;5:e12966.
101. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 2015;161:228–39.
102. Treiber CD, Waddell S. Resolving the prevalence of somatic transposition in *Drosophila*. eLife. 2017;6:e28297.
103. Faulkner GJ, Garcia-Perez JL. L1 mosaicism in mammals: extent, effects, and evolution. Trends Genet. 2017;33:802–16.
104. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010;141:1253–61.
105. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012;337:967–71.
106. Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science. 2014;345:1251343.
107. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26:745–55.
108. Tang Z, Steranka JP, Ma S, Grivainis M, Rodić N, Huang CRL, et al. Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proc Natl Acad Sci U S A. 2017;114:E733–40.
109. Schauer SN, Carreira PE, Shukla R, Gerhardt DJ, Gerdes P, Sanchez-Luque FJ, et al. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. 2018;28:639–53.
110. Rodić N, Steranka JP, Makohon-Moore A, Moyer A, Shen P, Sharma R, et al. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med. 2015;21:1060–4.
111. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17:415–24.
112. Gorbunova V, Boeke JD, Helfand SL, Sedivy JM. Sleeping dogs of the genome. Science. 2014;346:1187–8.
113. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9:397–405.
114. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. Science. 2016;351:aac7247.
115. Daniel C, Behm M, Öhman M. The role of Alu elements in the cis-regulation of RNA processing. Cell Mol Life Sci. 2015;72:4063–76.
116. Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J, et al. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. Aging Cell. 2013;12:247–56.
117. De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. Aging. 2013;5:867–83.
118. Van Meter M, Kashyap M, Rezazadeh S, Geneva AJ, Morello TD, Seluanov A, et al. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. Nat Commun. 2014;5:5011.
119. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mob DNA. 2016;7:24.
120. Hedges DJ, Deininger PL. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res Mol Mech Mutagen. 2007;616:46–59.
121. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. Cell. 2015;162:974–86.
122. Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. Nat Rev Immunol. 2016;16:207–19.
123. Roulois D, Yau HL, Singhania R, Wang Y, Danesh A, Shen SY, et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. Cell. 2015;162:961–73.
124. Crow YJ, Manel N. Aicardi–Goutières syndrome and the type I interferonopathies. Nat Rev Immunol. 2015;15:429–40.
125. Stetson DB, Ko JS, Heidmann T, Medzhitov R. Trex1 prevents cell-intrinsic initiation of autoimmunity. Cell. 2008;134:587–98.

126. Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, et al. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology. 2016;13:7.
127. Perron H, Jouvin-Marche E, Michel M, Ounanian-Paraz A, Camelo S, Dumon A, et al. Multiple sclerosis retrovirus particles and recombinant envelope trigger an abnormal immune response in vitro, by inducing polyclonal Vβ16 T-lymphocyte activation. Virology. 2001;287:321–32.
128. Li W, Lee M-H, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med. 2015;7:307ra153.
129. Downey RF, Sullivan FJ, Wang-Johanning F, Ambs S, Giles FJ, Glynn SA. Human endogenous retrovirus K and cancer: innocent bystander or tumorigenic accomplice? Int J Cancer. 2015;137:1249–57.
130. Kassiotis G, Stoye JP. Making a virtue of necessity: the pleiotropic role of human endogenous retroviruses in cancer. Philos Trans R Soc Lond B Biol Sci. 2017;372:20160277.
131. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007;8:241–59.
132. Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, et al. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect. 2016;22:312–23.
133. Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences. Curr Opin Genet Dev. 2018;49:34–42.
134. Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, et al. Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. Cell. 2016;166:102–14.
135. Kapitonov VV, Koonin EV. Evolution of the RAG1–RAG2 locus: both proteins came from the same transposon. Biol Direct. 2015;10:20.
136. Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. Curr Opin Virol. 2017;25:81–9.
137. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. Proc Natl Acad Sci U S A. 2015;112:E487–96.
138. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012;33:663–71.
139. Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, et al. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. Proc Natl Acad Sci U S A. 2017;114:E10991–1000.
140. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, et al. The neuronal gene arc encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. Cell. 2018;172:275–88.e18.
141. Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like Gag protein Arc1 binds RNA and traffics across synaptic boutons. Cell. 2018;172:262–74.e11.
142. Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, et al. Intronic Alus influence alternative splicing. PLoS Genet. 2008;4:e1000204.
143. Schmitz J, Brosius J. Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie. 2011;93:1928–34.
144. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Pérez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol Spectr. 2015;3:MDNA3-0061-2014.
145. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A. 2006;103:17608–13.
146. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000;24:363–7.
147. Kubiak MR, Makalowska I. Protein-coding genes' retrocopies and their functions. Viruses. 2017;9:80.
148. Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016;26:301–14.
149. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9:e1003470.
150. Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. 2014;21:423–5.
151. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516:405–9.
152. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nat Genet. 2015;48:44–52.
153. Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-nucleolin partnership regulates early development and ESC identity. Cell. 2018;174:391–405.e19.
154. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet. 2017;49:1502–10.
155. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, et al. Widespread establishment and regulatory impact of Alu exons in human genes. Proc Natl Acad Sci U S A. 2011;108:2837–42.
156. Johnson R, Guigó R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA. 2014;20:959–76.
157. Kelley DR, Hendrickson DG, Tenen D, Rinn JL. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol. 2014;15:537.
158. Lubelsky Y, Ulitsky I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature. 2018;555:107–11.
159. Piriyapongsa J, Mariño-Ramírez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. Genetics. 2007;176:1323–37.
160. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression. Trends Genet. 2012;28:616–23.
161. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441:87–90.
162. Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013;45:325–9.
163. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351:1083–7.
164. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, et al. Transposable elements are the primary source of novelty in primate gene regulation. Genome Res. 2017;27:1623–33.
165. Thompson PJ, Macfarlan TS, Lorincz MC. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. Mol Cell. 2016;62:766–76.
166. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013;9:e1003504.
167. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A. 2007;104:18613–8.
168. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008;18:1752–62.
169. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963–76.
170. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLOS Genet. 2017;13:e1006883.
171. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42:631–4.
172. Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, et al. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. Proc Natl Acad Sci U S A. 2018;115:E5526–35.
173. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148:335–48.
174. Lunyak VV, Prefontaine GG, Nunez E, Cramer T, Ju BG, Ohgi KA, et al. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science. 2007;317:248–51.

175. Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, et al. MIR retrotransposon sequences provide insulators to the human genome. Proc Natl Acad Sci U S A. 2015;112:E4428–37.
176. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430:471–6.
177. Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, et al. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLOS Genet. 2011;7:e1002301.
178. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.
179. Michaud EJ, Van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT, Woychik RP. Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. Genes Dev. 1994;8:1463–72.
180. Morgan HD, Sutherland HGE, Martin DIK, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999;23:314–8.
181. Ong-Abdullah M, Ordway JM, Jiang N, Ooi S-E, Kok S-Y, Sarpan N, et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature. 2015;525:533–7.
182. Davidson EH, Britten RJ. Regulation of gene expression: possible role of repetitive sequences. Science. 1979;204:1052–9.
183. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43:1154–9.
184. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 2015;10:551–61.
185. Notwell JH, Chung T, Heavner W, Bejerano G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat Commun. 2015;6:6644.
186. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015;11:e1004915.
187. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018. https://doi.org/10.1038/s41576-018-0050-x.
188. Niu D, Wei H-J, Lin L, George H, Wang T, Lee I-H, et al. Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. Science. 2017;357:1303–7.
189. Amabile A, Migliara A, Capasso P, Biffi M, Cittaro D, Naldini L, et al. Inheritable silencing of endogenous genes by hit-and-run targeted epigenetic editing. Cell. 2016;167:219–32.e14.
190. Guallar D, Bi X, Pardavila JA, Huang X, Saenz C, Shi X, et al. RNA-dependent chromatin targeting of TET2 for endogenous retrovirus control in pluripotent stem cells. Nat Genet. 2018;50:443–51.
191. Fuentes DR, Swigut T, Wysocka J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. Elife. 2018;7:e35989.
192. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
193. Wang J, Huda A, Lunyak VV, Jordan IK. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinformatics. 2010;26:2501–8.
194. Kahles A, Behr J, Rätsch G. MMR: a tool for read multi-mapper resolution. Bioinformatics. 2016;32:770–2.
195. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
196. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5:578–90.
197. Eddy SR. The ENCODE project: missteps overshadowing a success. Curr Biol. 2013;23:R259–61.
198. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A. 2013;110:5294–300.
199. Ohno S. So much "junk" DNA in our genome. In: Smith HH, editor. Evolution of genetic systems. Brookhaven: Symp Biol; 1972. p. 366–70.
200. Venuto D, Bourque G. Identifying co-opted transposable elements using comparative epigenomics. Dev Growth Differ. 2018;60:53–62.

A.7 Expression analysis

Recommended reading for Chapter 7 (listed in Section 7.9).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Wang et al. (2009) — RNA-seq, a revolutionary tool for transcriptomics; why sequencing replaced arrays.
  DOI · via SU library (SU login)

  RNA-Seq: a revolutionary tool for transcriptomics. Wang et al. (2009). *Nature Reviews Genetics*. doi:10.1038/nrg2484

  > RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

- Conesa et al. (2016) — a survey of best practices for RNA-seq data analysis, from design to interpretation.
  full text below · DOI · open access (CC BY 4.0)

  A survey of best practices for RNA-seq data analysis. Conesa et al. (2016). *Genome Biology*. doi:10.1186/s13059-016-0881-8

  > RNA-sequencing (RNA-seq) has a wide variety of applications, but no single analysis pipeline can be used in all cases. We review all of the major steps in RNA-seq data analysis, including experimental design, quality control, read alignment, quantification of gene and transcript levels, visualization, differential gene expression, alternative splicing, functional analysis, gene fusion detection and eQTL mapping. We highlight the challenges associated with each step. We discuss the analysis of small RNAs and the integration of RNA-seq with other functional genomics techniques. Finally, we discuss the outlook for novel technologies that are changing the state of the art in transcriptomics.

- Love et al. (2014) — DESeq2; negative binomial models, dispersion and fold-change shrinkage.
  full text below · DOI · open access (CC BY 4.0)

  Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Love et al. (2014). *Genome Biology*. doi:10.1186/s13059-014-0550-8

  > In comparative high-throughput sequencing assays, a fundamental task is the analysis of count data, such as read counts per gene in RNA-seq, for evidence of systematic changes across experimental conditions. Small replicate numbers, discreteness, large dynamic range and the presence of outliers require a suitable statistical approach. We present DESeq2, a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. This enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression. The DESeq2 package is available at http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html webcite.

- Benjamini and Hochberg (1995) — the false discovery rate; the most important idea in genome-scale statistics.
  DOI · via SU library (SU login)

  Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Benjamini & Hochberg (1995). *Journal of the Royal Statistical Society Series B: Statistical Methodology*. doi:10.1111/j.2517-6161.1995.tb02031.x

> SUMMARY The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses — the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferronitype procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

- Luecken and Theis (2019) — best practices in single-cell RNA-seq analysis, step by step.
  full text below · DOI · open access (CC BY 4.0)

  Current best practices in single-cell RNA-seq analysis: a tutorial. Luecken & Theis (2019). *Molecular Systems Biology*. doi:10.15252/msb.20188746

  > Single-cell RNA-seq has enabled gene expression to be studied at an unprecedented resolution. The promise of this technology is attracting a growing user base for single-cell analysis methods. As more analysis tools are becoming available, it is becoming increasingly difficult to navigate this landscape and produce an up-to-date workflow to analyse one's data. Here, we detail the steps of a typical single-cell RNA-seq analysis, including pre-processing (quality control, normalization, data correction, feature selection, and dimensionality reduction) and cell- and gene-level downstream analysis. We formulate current best-practice recommendations for these steps based on independent comparison studies. We have integrated these best-practice recommendations into a workflow, which we apply to a public dataset to further illustrate how these steps work in practice. Our documented case study can be found at https://www.github.com/theislab/single-cell-tutorial This review will serve as a workflow tutorial for new entrants into the field, and help established users update their analysis pipelines.

- Nam et al. (2020) — cancer evolution studied with single-cell multi-omics.
  DOI · via SU library (SU login)

  Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nam et al. (2020). *Nature Reviews Genetics*. doi:10.1038/s41576-020-0265-5

  > Cancer represents an evolutionary process through which growing malignant populations genetically diversify, leading to tumour progression, relapse and resistance to therapy. In addition to genetic diversity, the cell-to-cell variation that fuels evolutionary selection also manifests in cellular states, epigenetic profiles, spatial distributions and interactions with the microenvironment. Therefore, the study of cancer requires the integration of multiple heritable dimensions at the resolution of the single cell - the atomic unit of somatic evolution. In this Review, we discuss emerging analytic and experimental technologies for single-cell multi-omics that enable the capture and integration of multiple data modalities to inform the study of cancer evolution. These data show that cancer results from a complex interplay between genetic and non-genetic determinants of somatic evolution.

The open-access papers are reproduced below under their original licences.

Conesa et al. Genome Biology (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

Abstract

RNA-sequencing (RNA-seq) has a wide variety of applications, but no single analysis pipeline can be used in all cases. We review all of the major steps in RNA-seq data analysis, including experimental design, quality control, read alignment, quantification of gene and transcript levels, visualization, differential gene expression, alternative splicing, functional analysis, gene fusion detection and eQTL mapping. We highlight the challenges associated with each step. We discuss the analysis of small RNAs and the integration of RNA-seq with other functional genomics techniques. Finally, we discuss the outlook for novel technologies that are changing the state of the art in transcriptomics.

Background

Transcript identification and the quantification of gene expression have been distinct core activities in molecular biology ever since the discovery of RNA's role as the key intermediate between the genome and the proteome. The power of sequencing RNA lies in the fact that the twin aspects of discovery and quantification can be combined in a single high-throughput sequencing assay called RNA-sequencing (RNA-seq). The pervasive adoption of RNA-seq has spread well beyond the genomics community and has become a standard part of the toolkit used by the life sciences research community. Many variations of RNA-seq protocols and analyses have been published, making it challenging for new users to appreciate all of the steps necessary to conduct an RNA-seq study properly.

There is no optimal pipeline for the variety of different applications and analysis scenarios in which RNA-seq can be used. Scientists plan experiments and adopt different analysis strategies depending on the organism being studied and their research goals. For example, if a genome sequence is available for the studied organism, it should be possible to identify transcripts by mapping RNA-seq reads onto the genome. By contrast, for organisms without sequenced genomes, quantification would be achieved by first assembling reads de novo into contigs and then mapping these contigs onto the transcriptome. For well-annotated genomes such as the human genome, researchers may choose to base their RNA-seq analysis on the existing annotated reference transcriptome alone, or might try to identify new transcripts and their differential regulation. Furthermore, investigators might be interested only in messenger RNA isoform expression or microRNA (miRNA) levels or allele variant identification. Both the experimental design and the analysis procedures will vary greatly in each of these cases. RNA-seq can be used solo for transcriptome profiling or in combination with other functional genomics methods to enhance the analysis of gene expression. Finally, RNA-seq can be coupled with different types of biochemical assay to analyze many other aspects of RNA biology, such as RNA–protein binding, RNA structure, or RNA–RNA interactions. These applications are, however, beyond the scope of this review as we focus on 'typical' RNA-seq.

Every RNA-seq experimental scenario could potentially have different optimal methods for transcript quantification, normalization, and ultimately differential expression analysis. Moreover, quality control checks should be applied pertinently at different stages of the analysis to ensure both reproducibility and reliability of the results. Our focus is to outline current standards

* Correspondence: aconesa@ufl.edu; pm12@sanger.ac.uk; ali.mortazavi@uci.edu
[1]Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32603, USA
[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[16]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697-2300, USA
Full list of author information is available at the end of the article

BioMed Central

© 2016 Conesa et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

and resources for the bioinformatics analysis of RNA-seq data. We do not aim to provide an exhaustive compilation of resources or software tools nor to indicate one best analysis pipeline. Rather, we aim to provide a commented guideline for RNA-seq data analysis. Figure 1 depicts a generic roadmap for experimental design and analysis using standard Illumina sequencing. We also briefly list several data integration paradigms that have been proposed and comment on their potential and limitations. We finally discuss the opportunities as well as challenges provided by single-cell RNA-seq and long-read technologies when compared to traditional short-read RNA-seq.

Experimental design

A crucial prerequisite for a successful RNA-seq study is that the data generated have the potential to answer the biological questions of interest. This is achieved by first defining a good experimental design, that is, by choosing the library type, sequencing depth and number of replicates appropriate for the biological system under study, and second by planning an adequate execution of the sequencing experiment itself, ensuring that data acquisition does not become contaminated with unnecessary biases. In this section, we discuss both considerations.

One important aspect of the experimental design is the RNA-extraction protocol used to remove the highly abundant ribosomal RNA (rRNA), which typically constitutes over 90 % of total RNA in the cell, leaving the 1–2 % comprising messenger RNA (mRNA) that we are normally interested in. For eukaryotes, this involves choosing whether to enrich for mRNA using poly(A) selection or to deplete rRNA. Poly(A) selection typically requires a relatively high proportion of mRNA with minimal degradation as measured by RNA integrity number (RIN), which normally yields a higher overall fraction of reads falling onto known exons. Many biologically relevant samples (such as tissue biopsies) cannot, however, be obtained in great enough quantity or good enough mRNA integrity to produce good poly(A) RNA-seq libraries and therefore require ribosomal depletion. For bacterial samples, in which mRNA is not polyadenylated,

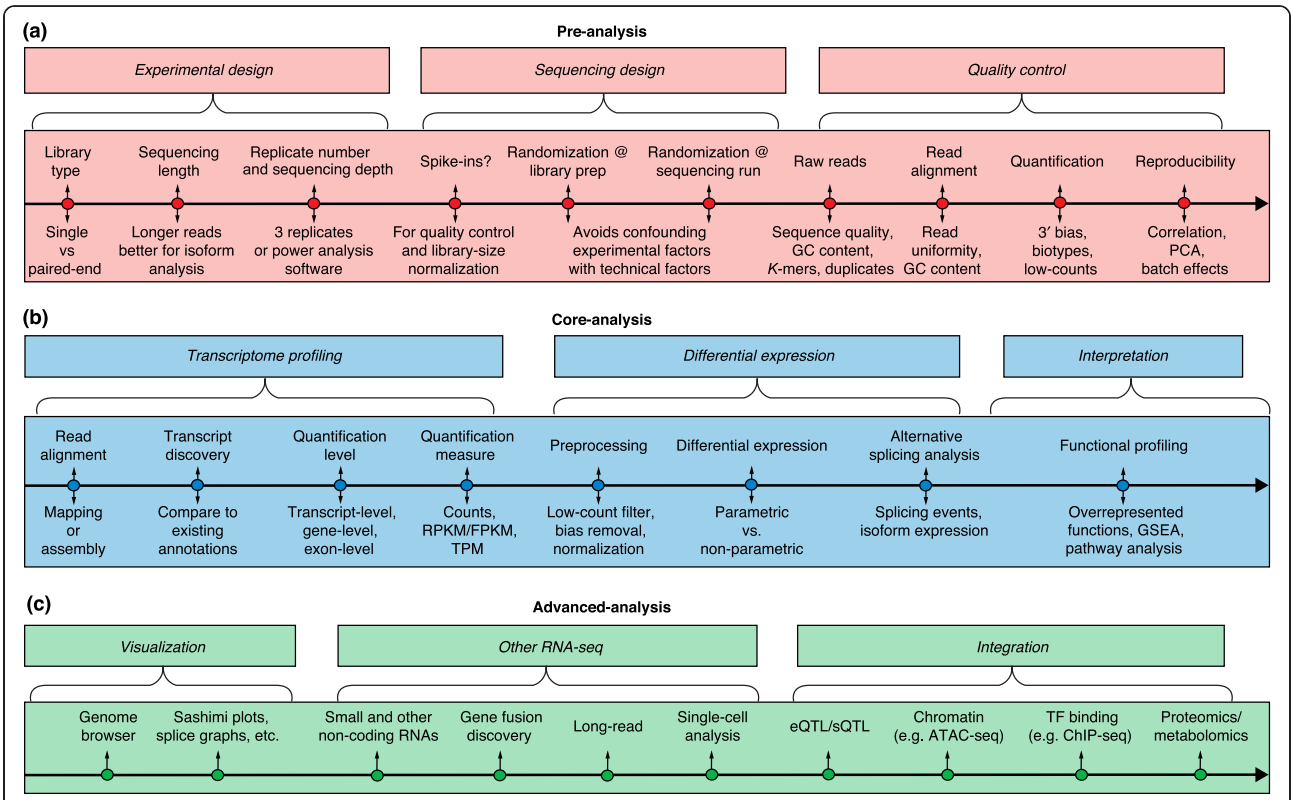


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

the only viable alternative is ribosomal depletion. Another consideration is whether to generate strand-preserving libraries. The first generation of Illumina-based RNA-seq used random hexamer priming to reverse-transcribe poly(A)-selected mRNA. This methodology did not retain information contained on the DNA strand that is actually expressed [1] and therefore complicates the analysis and quantification of antisense or overlapping transcripts. Several strand-specific protocols [2], such as the widely used dUTP method, extend the original protocol by incorporating UTP nucleotides during the second cDNA synthesis step, prior to adapter ligation followed by digestion of the strand containing dUTP [3]. In all cases, the size of the final fragments (usually less than 500 bp for Illumina) will be crucial for proper sequencing and subsequent analysis. Furthermore, sequencing can involve single-end (SE) or paired-end (PE) reads, although the latter is preferable for de novo transcript discovery or isoform expression analysis [4, 5]. Similarly, longer reads improve mappability and transcript identification [5, 6]. The best sequencing option depends on the analysis goals. The cheaper, short SE reads are normally sufficient for studies of gene expression levels in well-annotated organisms, whereas longer and PE reads are preferable to characterize poorly annotated transcriptomes.

Another important factor is sequencing depth or library size, which is the number of sequenced reads for a given sample. More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level [1]. Nevertheless, optimal sequencing depth again depends on the aims of the experiment. While some authors will argue that as few as five million mapped reads are sufficient to quantify accurately medium to highly expressed genes in most eukaryotic transcriptomes, others will sequence up to 100 million reads to quantify precisely genes and transcripts that have low expression levels [7]. When studying single cells, which have limited sample complexity, quantification is often carried out with just one million reads but may be done reliably for highly expressed genes with as few as 50,000 reads [8]; even 20,000 reads have been used to differentiate cell types in splenic tissue [9]. Moreover, optimal library size depends on the complexity of the targeted transcriptome. Experimental results suggest that deep sequencing improves quantification and identification but might also result in the detection of transcriptional noise and off-target transcripts [10]. Saturation curves can be used to assess the improvement in transcriptome coverage to be expected at a given sequencing depth [10].

Finally, a crucial design factor is the number of replicates. The number of replicates that should be included in a RNA-seq experiment depends on both the amount of technical variability in the RNA-seq procedures and the biological variability of the system under study, as well as on the desired statistical power (that is, the capacity for detecting statistically significant differences in gene expression between experimental groups). These two aspects are part of power analysis calculations (Fig. 1a; Box 1).

The adequate planning of sequencing experiments so as to avoid technical biases is as important as good

Box 1. Number of replicates

Three factors determine the number of replicates required in a RNA-seq experiment. The first factor is the variability in the measurements, which is influenced by the technical noise and the biological variation. While reproducibility in RNA-seq is usually high at the level of sequencing [1, 45], other steps such as RNA extraction and library preparation are noisier and may introduce biases in the data that can be minimized by adopting good experimental procedures (Box 2). Biological variability is particular to each experimental system and is harder to control [189]. Nevertheless, biological replication is required if inference on the population is to be made, with three replicates being the minimum for any inferential analysis. For a proper statistical power analysis, estimates of the within-group variance and gene expression levels are required. This information is typically not available beforehand but can be obtained from similar experiments. The exact power will depend on the method used for differential expression analysis, and software packages exist that provide a theoretical estimate of power over a range of variables, given the within-group variance of the samples, which is intrinsic to the experiment [190, 191]. Table 1 shows an example of statistical power calculations over a range of fold-changes (or effect sizes) and number of replicates in a human blood RNA-seq sample sequenced at 30 million mapped reads. It should be noted that these estimates apply to the average gene expression level, but as dynamic ranges in RNA-seq data are large, the probability that highly expressed genes will be detected as differentially expressed is greater than that for low-count genes [192]. For methods that return a false discovery rate (FDR), the proportion of genes that are highly expressed out of the total set of genes being tested will also influence the power of detection after multiple testing correction [193]. Filtering out genes that are expressed at low levels prior to differential expression analysis reduces the severity of the correction and may improve the power of detection [20]. Increasing sequencing depth also can improve statistical power for lowly expressed genes [10, 194], and for any given sample there exists a level of sequencing at which power improvement is best achieved by increasing the number of replicates [195]. Tools such as Scotty are available to calculate the best trade-off between sequencing depth and replicate number given some budgetary constraints [191].

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

experimental design, especially when the experiment involves a large number of samples that need to be processed in several batches. In this case, including controls, randomizing sample processing and smart management of sequencing runs are crucial to obtain error-free data (Fig. 1a; Box 2).

Analysis of the RNA-seq data

The actual analysis of RNA-seq data has as many variations as there are applications of the technology. In this section, we address all of the major analysis steps for a typical RNA-seq experiment, which involve quality control, read alignment with and without a reference genome, obtaining metrics for gene and transcript expression, and approaches for detecting differential gene expression. We also discuss analysis options for applications of RNA-seq involving alternative splicing, fusion transcripts and small RNA expression. Finally, we review useful packages for data visualization.

Box 2. Experiment execution choices

RNA-seq library preparation and sequencing procedures include a number of steps (RNA fragmentation, cDNA synthesis, adapter ligation, PCR amplification, bar-coding, and lane loading) that might introduce biases into the resulting data [196]. Including exogenous reference transcripts ('spike-ins') is useful both for quality control [1, 197] and for library-size normalization [198]. For bias minimization, we recommend following the suggestions made by Van Dijk et al. [199], such as the use of adapters with random nucleotides at the extremities or the use of chemical-based fragmentation instead of RNase III-based fragmentation. If the RNA-seq experiment is large and samples have to be processed in different batches and/or Illumina runs, caution should be taken to randomize samples across library preparation batches and lanes so as to avoid technical factors becoming confounded with experimental factors. Another option, when samples are individually barcoded and multiple Illumina lanes are needed to achieve the desired sequencing depth, is to include all samples in each lane, which would minimize any possible lane effect.

Quality-control checkpoints

The acquisition of RNA-seq data consists of several steps — obtaining raw reads, read alignment and quantification. At each of these steps, specific checks should be applied to monitor the quality of the data (Fig. 1a).

Raw reads

Quality control for the raw reads involves the analysis of sequence quality, GC content, the presence of adaptors, overrepresented *k*-mers and duplicated reads in order to detect sequencing errors, PCR artifacts or contaminations. Acceptable duplication, *k*-mer or GC content levels are experiment- and organism-specific, but these values should be homogeneous for samples in the same experiments. We recommend that outliers with over 30 % disagreement to be discarded. FastQC [11] is a popular tool to perform these analyses on Illumina reads, whereas NGSQC [12] can be applied to any platform. As a general rule, read quality decreases towards the 3' end of reads, and if it becomes too low, bases should be removed to improve mappability. Software tools such as the FASTX-Toolkit [13] and Trimmomatic [14] can be used to discard low-quality reads, trim adaptor sequences, and eliminate poor-quality bases.

Read alignment

Reads are typically mapped to either a genome or a transcriptome, as will be discussed later. An important mapping quality parameter is the percentage of mapped reads, which is a global indicator of the overall sequencing accuracy and of the presence of contaminating DNA. For example, we expect between 70 and 90 % of regular RNA-seq reads to map onto the human genome (depending on the read mapper used) [15], with a significant fraction of reads mapping to a limited number of identical regions equally well ('multi-mapping reads'). When reads are mapped against the transcriptome, we expect slightly lower total mapping percentages because reads coming from unannotated transcripts will be lost, and significantly more multi-mapping reads because of reads falling onto exons that are shared by different transcript isoforms of the same gene.

Other important parameters are the uniformity of read coverage on exons and the mapped strand. If reads

primarily accumulate at the 3' end of transcripts in poly(A)-selected samples, this might indicate low RNA quality in the starting material. The GC content of mapped reads may reveal PCR biases. Tools for quality control in mapping include Picard [16], RSeQC [17] and Qualimap [18].

Quantification

Once actual transcript quantification values have been calculated, they should be checked for GC content and gene length biases so that correcting normalization methods can be applied if necessary. If the reference transcriptome is well annotated, researchers could analyze the biotype composition of the sample, which is indicative of the quality of the RNA purification step. For example, rRNA and small RNAs should not be present in regular polyA longRNA preparations [10, 19]. A number of R packages (such as NOISeq [19] or EDASeq [20]) provide useful plots for quality control of count data.

Reproducibility

The quality-control steps described above involve individual samples. In addition, it is also crucial to assess the global quality of the RNA-seq dataset by checking on the reproducibility among replicates and for possible batch effects. Reproducibility among technical replicates should be generally high (Spearman $R^2 > 0.9$) [1], but no clear standard exists for biological replicates, as this depends on the heterogeneity of the experimental system. If gene expression differences exist among experimental conditions, it should be expected that biological replicates of the same condition will cluster together in a principal component analysis (PCA).

Transcript identification

When a reference genome is available, RNA-seq analysis will normally involve the mapping of the reads onto the reference genome or transcriptome to infer which transcripts are expressed. Mapping solely to the reference transcriptome of a known species precludes the discovery of new, unannotated transcripts and focuses the analysis on quantification alone. By contrast, if the organism does not have a sequenced genome, then the analysis path is first to assemble reads into longer contigs and then to treat these contigs as the expressed transcriptome to which reads are mapped back again for quantification. In either case, read coverage can be used to quantify transcript expression level (Fig. 1b). A basic choice is whether transcript identification and quantification are done sequentially or simultaneously.

Alignment

Two alternatives are possible when a reference sequence is available: mapping to the genome or mapping to the annotated transcriptome (Fig. 2a, b; Box 3). Regardless

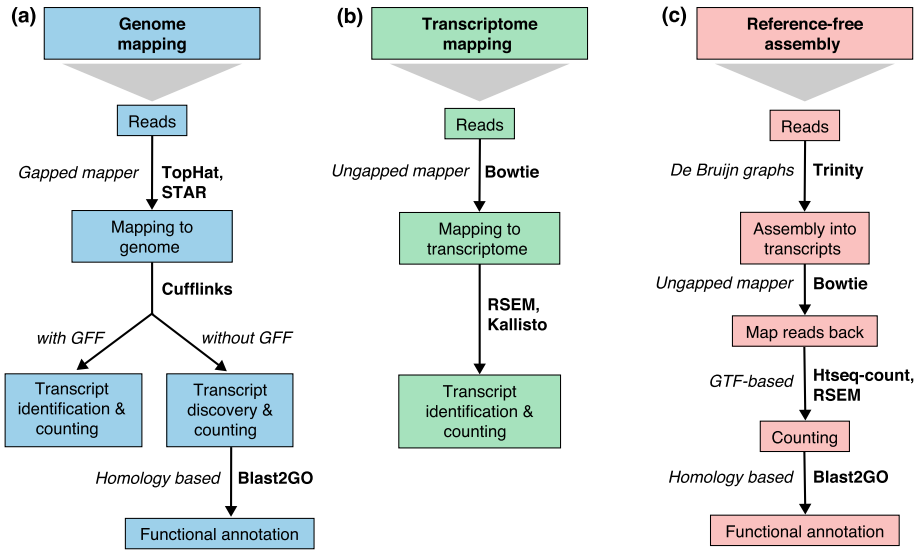


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in (**b**) followed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used at each analysis step are indicated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

Box 3. Mapping to a reference

Mapping to a reference genome allows for the identification of novel genes or transcripts, and requires the use of a gapped or spliced mapper as reads may span splice junctions. The challenge is to identify splice junctions correctly, especially when sequencing errors or differences with the reference exist or when non-canonical junctions and fusion transcripts are sought. One of the most popular RNA-seq mappers, TopHat, follows a two-step strategy in which unspliced reads are first mapped to locate exons, then unmapped reads are split and aligned independently to identify exon junctions [200, 201]. Several other mappers exist that are optimized to identify SNPs or indels (GSNAP [202], PALMapper [203] MapSplice [204]), detect non-canonical splice junctions (STAR [15], MapSplice [204]), achieve ultra-fast mapping (GEM [205]) or map long-reads (STAR [15]). Important parameters to consider during mapping are the strandedness of the RNA-seq library, the number of mismatches to accept, the length and type of reads (SE or PE), and the length of sequenced fragments. In addition, existing gene models can be leveraged by supplying an annotation file to some read mapper in order to map exon coordinates accurately and to help in identifying splicing events. The choice of gene model can also have a strong impact on the quantification and differential expression analysis [206]. We refer the reader to [30] for a comprehensive comparison of RNA-seq mappers. If the transcriptome annotation is comprehensive (for example, in mouse or human), researchers may choose to map directly to a Fasta-format file of all transcript sequences for all genes of interests. In this case, no gapped alignment is needed and unspliced mappers such as Bowtie [207] can be used (Fig. 2b). Mapping to the transcriptome is generally faster but does not allow de novo transcript discovery.

of whether a genome or transcriptome reference is used, reads may map uniquely (they can be assigned to only one position in the reference) or could be multi-mapped reads (multireads). Genomic multireads are primarily due to repetitive sequences or shared domains of paralogous genes. They normally account for a significant fraction of the mapping output when mapped onto the genome and should not be discarded. When the reference is the transcriptome, multi-mapping arises even more often because a read that would have been uniquely mapped on the genome would map equally well to all gene isoforms in the transcriptome that share the exon. In either case — genome or transcriptome mapping — transcript identification and quantification become important challenges for alternatively expressed genes.

Transcript discovery

Identifying novel transcripts using the short reads provided by Illumina technology is one of the most challenging tasks in RNA-seq. Short reads rarely span across several splice junctions and thus make it difficult to directly infer all full-length transcripts. In addition, it is difficult to identify transcription start and end sites [21], and tools such as GRIT [22] that incorporate other data such as 5' ends from CAGE or RAMPAGE typically have a better chance of annotating the major expressed isoforms correctly. In any case, PE reads and higher coverage help to reconstruct lowly expressed transcripts, and replicates are essential to resolve false-positive calls (that is, mapping artifacts or contaminations) at the low end of signal detection. Several methods, such as Cufflinks [23], iReckon [24], SLIDE [25] and StringTie [26], incorporate existing annotations by adding them to the possible list of isoforms. Montebello [27] couples isoform discovery and quantification using a likelihood-based Monte Carlo algorithm to boost performance. Gene-finding tools such as Augustus [28] can incorporate RNA-seq data to better annotate protein-coding transcripts, but perform worse on non-coding transcripts [29]. In general, accurate transcript reconstruction from short reads is difficult, and methods typically show substantial disagreement [29].

De novo transcript reconstruction

When a reference genome is not available or is incomplete, RNA-seq reads can be assembled de novo (Fig. 2c) into a transcriptome using packages such as SOAPdenovo-Trans [30], Oases [31], Trans-ABySS [32] or Trinity [33]. In general, PE strand-specific sequencing and long reads are preferred because they are more informative [33]. Although it is impossible to assemble lowly expressed transcripts that lack enough coverage for a reliable assembly, too many reads are also problematic because they lead to potential misassembly and increased runtimes. Therefore, in silico reduction of the number of reads is recommended for deeply sequenced samples [33]. For comparative analyses across samples, it is advisable to combine all reads from multiple samples into a single input in order to obtain a consolidated set of contigs (transcripts), followed by mapping back of the short reads for expression estimation [33].

Either with a reference or de novo, the complete reconstruction of transcriptomes using short-read Illumina technology remains a challenging problem, and in many cases de novo assembly results in tens or hundreds of contigs accounting for fragmented transcripts. Emerging long-read technologies, such as SMRT from Pacific Biosciences, provide reads that are long enough to sequence complete transcripts for most genes and are a promising alternative that is discussed further in the "Outlook" section below.

Transcript quantification

The most common application of RNA-seq is to estimate gene and transcript expression. This application is primarily based on the number of reads that map to each transcript sequence, although there are algorithms such as Sailfish that rely on *k*-mer counting in reads without the need for mapping [34]. The simplest approach to quantification is to aggregate raw counts of mapped reads using programs such as HTSeq-count [35] or featureCounts [36]. This gene-level (rather than transcript-level) quantification approach utilizes a gene transfer format (GTF) file [37] containing the genome coordinates of exons and genes, and often discard multi-reads. Raw read counts alone are not sufficient to compare expression levels among samples, as these values are affected by factors such as transcript length, total number of reads, and sequencing biases. The measure RPKM (reads per kilobase of exon model per million reads) [1] is a within-sample normalization method that will remove the feature-length and library-size effects. This measure and its subsequent derivatives FPKM (fragments per kilobase of exon model per million mapped reads), a within-sample normalized transcript expression measure analogous to RPKs, and TPM (transcripts per million) are the most frequently reported RNA-seq gene expression values. It should be noted that RPKM and FPKM are equivalent for SE reads and that FPKM can be converted into TPM using a simple formula [38]. The dichotomy of within-sample and between-sample comparisons has led to a lot of confusion in the literature. Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples, but it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads. Furthermore, programs such as Cufflinks that estimate gene length from the data can find significant differences in gene length between samples that cannot be ignored. TPMs, which effectively normalize for the differences in composition of the transcripts in the denominator rather than simply dividing by the number of reads in the library, are considered more comparable between samples of different origins and composition but can still suffer some biases. These must be addressed with normalization techniques such as TMM.

Several sophisticated algorithms have been developed to estimate transcript-level expression by tackling the problem of related transcripts' sharing most of their reads. Cufflinks [39] estimates transcript expression from a mapping to the genome obtained from mappers such as TopHat using an expectation-maximization approach that estimates transcript abundances. This approach takes into account biases such as the non-uniform read distribution along the gene length. Cufflinks was designed to take advantage of PE reads, and may use GTF information to identify expressed transcripts, or can infer transcripts de novo from the mapping data alone. Algorithms that quantify expression from transcriptome mappings include RSEM (RNA-Seq by Expectation Maximization) [40], eXpress [41], Sailfish [35] and kallisto [42] among others. These methods allocate multi-mapping reads among transcript and output within-sample normalized values corrected for sequencing biases [35, 41, 43]. Additionally, the RSEM algorithm uses an expectation maximization approach that returns TPM values [40]. NURD [44] provides an efficient way of estimating transcript expression from SE reads with a low memory and computing cost.

Differential gene expression analysis

Differential expression analysis (Fig. 1b) requires that gene expression values should be compared among samples. RPKM, FPKM, and TPM normalize away the most important factor for comparing samples, which is sequencing depth, whether directly or by accounting for the number of transcripts, which can differ significantly between samples. These approaches rely on normalizing methods that are based on total or effective counts, and tend to perform poorly when samples have heterogeneous transcript distributions, that is, when highly and differentially expressed features can skew the count distribution [45, 46]. Normalization methods that take this into account are TMM [47], DESeq [48], PoissonSeq [49] and UpperQuartile [45], which ignore highly variable and/or highly expressed features. Additional factors that interfere with intra-sample comparisons include changes in transcript length across samples or conditions [50], positional biases in coverage along the transcript (which are accounted for in Cufflinks), average fragment size [43], and the GC contents of genes (corrected in the EDAseq package [21]). The NOISeq R package [20] contains a wide variety of diagnostic plots to identify sources of biases in RNA-seq data and to apply appropriate normalization procedures in each case. Finally, despite these sample-specific normalization methods, batch effects may still be present in the data. These effects can be minimized by appropriate experimental design [51] or, alternatively, removed by batch-correction methods such as COMBAT [52] or ARSyN [20, 53]. These approaches, although initially developed for microarray data, have been shown to work well with normalized RNA-seq data (STATegra project, unpublished).

As RNA-seq quantification is based on read counts that are absolutely or probabilistically assigned to transcripts, the first approaches to compute differential expression used discrete probability distributions, such as

the Poisson or negative binomial [48, 54]. The negative binomial distribution (also known as the gamma-Poisson distribution) is a generalization of the Poisson distribution, allowing for additional variance (called overdispersion) beyond the variance expected from randomly sampling from a pool of molecules that are characteristic of RNA-seq data. However, the use of discrete distributions is not required for accurate analysis of differential expression as long as the sampling variance of small read counts is taken into account (most important for experiments with small numbers of replicates). Methods for transforming normalized counts of RNA-seq reads while learning the variance structure of the data have been shown to perform well in comparison to the discrete distribution approaches described above [55, 56]. Moreover, after extensive normalization (including TMM and batch removal), the data might have lost their discrete nature and be more akin to a continuous distribution.

Some methods, such as the popular edgeR [57], take as input raw read counts and introduce possible bias sources into the statistical model to perform an integrated normalization as well as a differential expression analysis. In other methods, the differential expression requires the data to be previously normalized to remove all possible biases. DESeq2, like edgeR, uses the negative binomial as the reference distribution and provides its own normalization approach [48, 58]. baySeq [59] and EBSeq [60] are Bayesian approaches, also based on the negative binomial model, that define a collection of models to describe the differences among experimental groups and to compute the posterior probability of each one of them for each gene. Other approaches include data transformation methods that take into account the sampling variance of small read counts and create discrete gene expression distributions that can be analyzed by regular linear models [55]. Finally, non-parametric approaches such as NOISeq [10] or SAMseq [61] make minimal assumptions about the data and estimate the null distribution for inferential analysis from the actual data alone. For small-scale studies that compare two samples with no or few replicates, the estimation of the negative binomial distribution can be noisy. In such cases, simpler methods based on the Poisson distribution, such as DEGseq [62], or on empirical distributions (NOISeq [10]) can be an alternative, although it should be strongly stressed that, in the absence of biological replication, no population inference can be made and hence any p value calculation is invalid. Methods that analyze RNA-seq data without replicates therefore only have exploratory value. Considering the drop in price of sequencing, we recommend that RNA-seq experiments have a minimum of three biological replicates when sample availability is not limiting to allow all of the differential expression methods to leverage reproducibility between replicates.

Recent independent comparison studies have demonstrated that the choice of the method (or even the version of a software package) can markedly affect the outcome of the analysis and that no single method is likely to perform favorably for all datasets [56, 63, 64] (Box 4). We therefore recommend thoroughly documenting the settings and version numbers of programs used and considering the repetition of important analyses using more than one package.

Alternative splicing analysis

Transcript-level differential expression analysis can potentially detect changes in the expression of transcript isoforms from the same gene, and specific algorithms for alternative splicing-focused analysis using RNA-seq have been proposed. These methods fall into two major categories. The first approach integrates isoform expression estimation with the detection of differential expression to reveal changes in the proportion of each isoform within the total gene expression. One such early method, BASIS, used a hierarchical Bayesian model to directly infer differentially expressed transcript isoforms [65]. CuffDiff2 estimates isoform expression first and then compares their differences. By integrating the two steps, the uncertainty in the first step is taken into consideration when performing the statistical analysis to look for differential isoform expression [66]. The flow difference metric (FDM) uses aligned cumulative transcript graphs from mapped exon reads and junction reads to infer isoforms and the Jensen-Shannon divergence to measure the difference [67]. Recently, Shi and Jiang [68] proposed a new method, rSeqDiff, that uses a hierarchical likelihood ratio test to detect differential gene expression without splicing change and differential isoform expression simultaneously. All these approaches are generally hampered by the intrinsic limitations of short-read sequencing for accurate identification at the isoform level, as discussed in the RNA-seq Genome Annotation Assessment Project paper [30].

The so-called 'exon-based' approach skips the estimation of isoform expression and detects signals of alternative splicing by comparing the distributions of reads on exons and junctions of the genes between the compared samples. This approach is based on the premise that differences in isoform expression can be tracked in the signals of exons and their junctions. DEXseq [69] and DSGSeq [70] adopt a similar idea to detect differentially spliced genes by testing for significant differences in read counts on exons (and junctions) of the genes. rMATS detects differential usage of exons by comparing exon-inclusion levels defined with junction reads [71]. rDiff detects differential isoform expression by comparing

Box 4. Comparison of software tools for detecting differential gene and transcript expression

Many statistical methods are available for detecting differential gene or transcript expression from RNA-seq data, and a major practical challenge is how to choose the most suitable tool for a particular data analysis job. Most comparison studies have focused on simulated datasets [56, 208, 209] or on samples to which exogenous RNA ('spike-in') has been added in known quantities [63, 196]. This enables a direct assessment of the sensitivity and specificity of the methods as well as their FDR control. As simulations typically rely on specific statistical distributions or on limited experimental datasets and as spike-in datasets represent only technical replicates with minimal variation, comparisons using simulated datasets have been complemented with more practical comparisons in real datasets with true biological replicates [64, 210, 211].

As yet, no clear consensus has been reached regarding the best practices and the field is continuing to evolve rapidly. However, some common findings have been made in multiple comparison studies and in different study settings. First, specific caution is needed with all the methods when the number of replicate samples is very small or for genes that are expressed at very low levels [55, 64, 209]. Among the tools, limma has been shown to perform well under many circumstances and it is also the fastest to run [56, 63, 64]. DESeq and edgeR perform similarly in ranking genes but are often relatively conservative or too liberal, respectively, in controlling FDR [63, 209, 210]. SAMseq performs well in terms of FDR but presents an acceptable sensitivity when the number of replicates is relatively high, at least 10 [20, 55, 209]. NOISeq and NOISeqBIO (the adaptation of NOISeq for biological replication) are more efficient in avoiding false positive calls at the cost of some sensitivity but perform well with different numbers of replicates [10, 20, 212]. Cuffdiff and Cuffdiff2 have performed surprisingly poorly in the comparisons [56, 63]. This probably reflects the fact that detecting differential expression at the transcript level remains challenging and involves uncertainties in assigning the reads to alternative isoforms. In a recent comparison, BitSeq compared favorably to other transcript-level packages such as Cuffdiff2 [196]. Besides the actual performance, other issues affecting the choice of the tool include ease of installation and use, computational requirements, and quality of documentation and instructions. Finally, an important consideration when choosing an analysis method is the experimental design. While some of the differential expression tools can only perform a pair-wise comparison, others such as edgeR [57], limma-voom [55], DESeq [48], DESeq2 [58], and maSigPro [213] can perform multiple comparisons, include different covariates or analyze time-series data.

read counts on alternative regions of the gene, either with or without annotated alternative isoforms [72]. DiffSplice uses alignment graphs to identify alternative splicing modules (ASMs) and identifies differential splicing using signals of the ASMs [73]. The advantage of exon or junction methods is their greater accuracy in identifying individual alternative splicing events. Exon-based methods are appropriate if the focus of the study is not on whole isoforms but on the inclusion and exclusion of specific exons and the functional protein domains (or regulatory features, in case of untranslated region exons) that they contain.

Visualization

Visualization of RNA-seq data (Fig. 1c) is, in general terms, similar to that of any other type of genomic sequencing data, and it can be done at the level of reads (using ReadXplorer [74], for example) or at the level of processed coverage (read pileup), unnormalized (for example, total count) or normalized, using genome browsers such as the UCSC browser [75], Integrative Genomics Viewer (IGV) [76] (Figure S1a in Additional file 1), Genome Maps [77], or Savant [78]. Some visualization tools are specifically designed for visualizing multiple RNA-seq samples, such as RNAseqViewer [79], which provides flexible ways to display the read abundances on exons, transcripts and junctions. Introns can be hidden to better display signals on the exons, and the heatmaps can help the visual comparison of signals on multiple samples (Figure S1b, c in Additional file 1). However, RNAseqViewer is slower than IGV.

Some of the software packages for differential gene expression analysis (such as DESeq2 or DEXseq in Bioconductor) have functions to enable the visualization of results, whereas others have been developed for visualization-exclusive purposes, such as CummeRbund (for CuffDiff [66]) or Sashimi plots, which can be used to visualize differentially spliced exons [80]. The advantage of Sashimi plots is that their display of junction reads is more intuitive and aesthetically pleasing when the number of samples is small (Figure S1d in Additional file 1). Sashimi, structure, and hive plots for splicing quantitative trait loci (sQTL) can be obtained using SplicePlot [81]. Splice graphs can be produced using SpliceSeq [82], and SplicingViewer [83] plots splice junctions and alternative splicing events. TraV [84] is a visualization tool that integrates data analysis, but its analytical methods are not applicable to large genomes.

Owing to the complexity of transcriptomes, efficient display of multiple layers of information is still a challenge. All of the tools are evolving rapidly and we can expect more comprehensive tools with desirable features to be available soon. Nevertheless, the existing tools are of great value for exploring results for individual genes

of biological interest to assess whether particular analyses' results can withstand detailed scrutiny or to reveal potential complications caused by artifacts, such as 3' biases or complicated transcript structures. Users should visualize changes in read coverage for genes that are deemed important or interesting on the basis of their analysis results to evaluate the robustness of their conclusions.

Gene fusion discovery

The discovery of fused genes that can arise from chromosomal rearrangements is analogous to novel isoform discovery, with the added challenge of a much larger search space as we can no longer assume that the transcript segments are co-linear on a single chromosome. Artifacts are common even using state-of-the-art tools, which necessitates post-processing using heuristic filters [85]. Artifacts primarily result from misalignment of read sequences due to polymorphisms, homology, and sequencing errors. Families of homologous genes, and highly polymorphic genes such as the HLA genes, produce reads that cannot be easily mapped uniquely to their location of origin in the reference genome. For genes with very high expression, the small but non-negligible sequencing error rate of RNA-seq will produce reads that map incorrectly to homologous loci. Filtering highly polymorphic genes and pairs of homologous genes is recommended [86, 87]. Also recommended is the filtering of highly expressed genes that are unlikely to be involved in gene fusions, such as ribosomal RNA [86]. Finally, a low ratio of chimeric to wild-type reads in the vicinity of the fusion boundary may indicate spurious mis-mapping of reads from a highly expressed gene (the transcript allele fraction described by Yoshihara et al. [87]).

Given successful prediction of chimeric sequences, the next step is the prioritization of gene fusions that have biological impact over more expected forms of genomic variation. Examples of expected variation include immunoglobulin (IG) rearrangements in tumor samples infiltrated by immune cells, transiently expressed transposons and nuclear mitochondrial DNA, and read-through chimeras produced by co-transcription of adjacent genes [88]. Care must be taken with filtering in order not to lose events of interest. For example, removing all fusions involving an IG gene may remove real IG fusions in lymphomas and other blood disorders; filtering fusions for which both genes are from the IG locus is preferred [88]. Transiently expressed genomic breakpoint sequences that are associated with real gene fusions often overlap transposons; these should be filtered unless they are associated with additional fusion isoforms from the same gene pair [89]. Read-through chimeras are easily identified as predictions involving alternative splicing between adjacent genes. Where possible, fusions should be filtered by their presence in a set of control datasets [87]. When control datasets are not available, artifacts can be identified by their presence in a large number of unrelated datasets, after excluding the possibility that they represent true recurrent fusions [90, 91].

Strong fusion-sequence predictions are characterized by distinct subsequences that each align with high specificity to one of the fused genes. As alignment specificity is highly correlated with sequence length, a strong prediction sequence is longer, with longer subsequences from each gene. Longer reads and larger insert sizes produce longer predicted sequences; thus, we recommend PE RNA-seq data with larger insert size over SE datasets or datasets with short insert size. Another indicator of prediction strength is splicing. For most known fusions, the genomic breakpoint is located in an intron of each gene [92] and the fusion boundary coincides with a splice site within each gene. Furthermore, fusion isoforms generally follow the splicing patterns of wild-type genes. Thus, high confidence predictions have fusion boundaries coincident with exon boundaries and exons matching wild-type exons [91]. Fusion discovery tools often incorporate some of the aforementioned ideas to rank fusion predictions [93, 94], though most studies apply additional custom heuristic filters to produce a list of high-quality fusion candidates [90, 91, 95].

Small RNAs

Next-generation sequencing represents an increasingly popular method to address questions concerning the biological roles of small RNAs (sRNAs). sRNAs are usually 18–34 nucleotides in length, and they include miRNAs, short-interfering RNAs (siRNAs), PIWI-interacting RNAs (piRNAs), and other classes of regulatory molecules. sRNA-seq libraries are rarely sequenced as deeply as regular RNA-seq libraries because of a lack of complexity, with a typical range of 2–10 million reads. Bioinformatics analysis of sRNA-seq data differs from standard RNA-seq protocols (Fig. 1c). Ligated adaptor sequences are first trimmed and the resulting read-length distribution is computed. In animals, there are usually peaks for 22 and 23 nucleotides, whereas in plants there are peaks for 21- and 24-nucleotide redundant reads. For instance, miRTools 2.0 [96], a tool for prediction and profiling of sRNA species, uses by default reads that are 18–30 bases long. The threshold value depends on the application, and in case of miRNAs is usually in the range of 19–25 nucleotides.

As in standard RNA-seq, sRNA reads must then be aligned to a reference genome or transcriptome sequences using standard tools, such as Bowtie2 [97], STAR [15], or Burrows-Wheeler Aligner (BWA) [98].

There are, however, some aligners (such as PatMaN [99] and MicroRazerS [100]) that have been designed to map short sequences with preset parameter value ranges suited for optimal alignment of short reads. The mapping itself may be performed with or without mismatches, the latter being used more commonly. In addition, reads that map beyond a predetermined set number of locations may be removed as putatively originating from repetitive elements. In the case of miRNAs, usually 5–20 distinct mappings per genome are allowed. sRNA reads are then simply counted to obtain expression values. However, users should also verify that their sRNA reads are not significantly contaminated by degraded mRNA, for example, by checking whether a miRNA library shows unexpected read coverage over the body of highly expressed genes such as *GAPDH* or *ACTB*.

Further analysis steps include comparison with known sRNAs and de novo identification of sRNAs. There are class-specific tools for this purpose, such as miRDeep [101] and miRDeep-P [102] for animal and plant miRNAs, respectively, or the trans-acting siRNA prediction tool at the UEA sRNA Workbench [103]. Tools such as miRTools 2.0 [96], ShortStack [104], and iMir [105] also exist for comprehensive annotation of sRNA libraries and for identification of diverse classes of sRNAs.

Functional profiling with RNA-seq

The last step in a standard transcriptomics study (Fig. 1b) is often the characterization of the molecular functions or pathways in which differentially expressed genes (DEGs) are involved. The two main approaches to functional characterization that were developed first for microarray technology are (a) comparing a list of DEGs against the rest of the genome for overrepresented functions, and (b) gene set enrichment analysis (GSEA), which is based on ranking the transcriptome according to a measurement of differential expression. RNA-seq biases such as gene length complicate the direct applications of these methods for count data and hence RNA-seq-specific tools have been proposed. For example, GOseq [106] estimates a bias effect (such as gene length) on differential expression results and adapts the traditional hypergeometric statistic used in the functional enrichment test to account for this bias. Similarly, the Gene Set Variation Analysis (GSVA) [107] or SeqGSEA [108] packages also combine splicing and implement enrichment analyses similar to GSEA.

Functional analysis requires the availability of sufficient functional annotation data for the transcriptome under study. Resources such as Gene Ontology [109], Bioconductor [110], DAVID [111, 112] or Babelomics [113] contain annotation data for most model species. However, novel transcripts discovered during de novo transcriptome assembly or reconstruction would lack at least some functional information and therefore annotation is necessary for functional profiling of those results. Protein-coding transcripts can be functionally annotated using orthology by searching for similar sequences in protein databases such as SwissProt [114] and in databases that contain conserved protein domains such as Pfam [115] and InterPro [116]. The use of standard vocabularies such as the Gene Ontology (GO) allows for some exchangeability of functional information across orthologs. Popular tools such as Blast2GO [117] allow massive annotation of complete transcriptome datasets against a variety of databases and controlled vocabularies. Typically, between 50 and 80 % of the transcripts reconstructed from RNA-seq data can be annotated with functional terms in this way. However, RNA-seq data also reveal that an important fraction of the transcriptome is lacking protein-coding potential. The functional annotation of these long non-coding RNAs is more challenging as their conservation is often less pronounced than that of protein-coding genes. The Rfam database [118] contains most well-characterized RNA families, such as ribosomal or transfer RNAs, while mirBase [119] or Miranda [120] are specialized in miRNAs. These resources can be used for similarity-based annotation of short non-coding RNAs, but no standard functional annotation procedures are available yet for other RNA types such as the long non-coding RNAs.

Integration with other data types

The integration of RNA-seq data with other types of genome-wide data (Fig. 1c) allows us to connect the regulation of gene expression with specific aspects of molecular physiology and functional genomics. Integrative analyses that incorporate RNA-seq data as the primary gene expression readout that is compared with other genomic experiments are becoming increasingly prevalent. Below, we discuss some of the additional challenges posed by such analyses.

DNA sequencing

The combination of RNA and DNA sequencing can be used for several purposes, such as single nucleotide polymorphism (SNP) discovery, RNA-editing analyses, or expression quantitative trait loci (eQTL) mapping. In a typical eQTL experiment, genotype and transcriptome profiles are obtained from the same tissue type across a relatively large number of individuals (>50) and correlations between genotype and expression levels are then detected. These associations can unravel the genetic basis of complex traits such as height [121], disease susceptibility [122] or even features of genome architecture [123, 124]. Large eQTL studies have shown that genetic variation affects the expression of most genes [125–128].

RNA-seq has two major advantages over array-based technologies for detecting eQTLs. First, it can identify variants that affect transcript processing. Second, reads that overlap heterozygous SNPs can be mapped to maternal and paternal chromosomes, enabling quantification of allele-specific expression within an individual [129]. Allele-specific signals provide additional information about a genetic effect on transcription, and a number of computational methods have recently become available that leverage these signals to boost power for association mapping [130–132]. One challenge of this approach is the computational burden, as billions of gene–SNP associations need to be tested; bootstrapping or permutation-based approaches [133] are frequently used [134, 135]. Many studies have focused on testing only SNPs in the *cis* region surrounding the gene in question, and computationally efficient approaches have been developed recently to allow extremely swift mapping of eQTLs genome-wide [136]. Moreover, the combination of RNA-seq and re-sequencing can be used both to remove false positives when inferring fusion genes [88] and to analyze copy number alterations [137].

DNA methylation

Pairwise DNA-methylation and RNA-seq integration, for the most part, has consisted of the analysis of the correlation between DEGs and methylation patterns [138–140]. General linear models [141–143], logistic regression models [143] and empirical Bayes model [144] have been attempted among other modeling approaches. The statistically significant correlations that were observed, however, accounted for relatively small effects. An interesting shift away from focusing on individual gene–CpG methylation correlations is to use a network-interaction-based approach to analyze RNA-seq in relation to DNA methylation. This approach identifies one or more sets of genes (also called modules) that have coordinated differential expression and differential methylation [145].

Chromatin features

The combination of RNA-seq and transcription factor (TF) chromatin immunoprecipitation sequencing (ChIP-seq) data can be used to remove false positives in ChIP-seq analysis and to suggest the activating or repressive effect of a TF on its target genes. For example, BETA [146] uses differential gene expression in combination with peaks from ChIP-seq experiments to call TF targets. In addition, ChIP-seq experiments involving histone modifications have been used to understand the general role of these epigenomic changes on gene expression [147, 148]. Other RNA-ChIP-sequencing integrative approaches are reviewed in [149]. Integration of open chromatin data such as that from FAIRE-seq and DNase-seq with RNA-seq has mostly been limited to verifying the expression status of genes that overlap a region of interest [150]. DNase-seq can be used for genome-wide footprinting of DNA-binding factors, and this in combination with the actual expression of genes can be used to infer active transcriptional networks [150].

MicroRNAs

Integration of RNA-seq and miRNA-seq data has the potential to unravel the regulatory effects of miRNAs on transcript steady-state levels. This analysis is challenging, however, because of the very noisy nature of miRNA target predictions, which hampers analyses based on correlations between miRNAs and their target genes. Associations might be found in databases such as mirWalk [151] and miRBase [152] that offer target prediction according to various algorithms. Tools such as CORNA [153], MMIA [154, 155], MAGIA [156], and SePIA [157] refine predictions by testing for significant associations between genes, miRNAs, pathways and GO terms, or by testing the relatedness or anticorrelation of the expression profiles of both the target genes and the associated miRNAs. In general, we recommend using miRNA–mRNA associations that are predicted by several algorithms. For example, in mouse, we found that requiring miRNA–mRNA association in five databases resulted in about 50 target mRNA predictions per miRNA (STATegra observations).

Proteomics and metabolomics

Integration of RNA-seq with proteomics is controversial because the two measurements show generally low correlation (~0.40 [158, 159]). Nevertheless, pairwise integration of proteomics and RNA-seq can be used to identify novel isoforms. Unreported peptides can be predicted from RNA-seq data and then used to complement databases normally queried in mass spectrometry as done by Low et al. [160]. Furthermore, post-translational editing events may be identified if peptides that are present in the mass spectrometry analysis are absent from the expressed genes of the RNA-seq dataset. Integration of transcriptomics with metabolomics data has been used to identify pathways that are regulated at both the gene expression and the metabolite level, and tools are available that visualize results within the pathway context (MassTRIX [161], Paintomics [162], VANTED v2 [163], and SteinerNet [164]).

Integration and visualization of multiple data types

Integration of more than two genomic data types is still at its infancy and not yet extensively applied to functional sequencing techniques, but there are already some tools that combine several data types. SNMNMF [165] and PIMiM [166] combine mRNA and miRNA expression

data with protein–protein, DNA–protein, and miRNA–mRNA interaction networks to identify miRNA–gene regulatory modules. MONA [167] combines different levels of functional genomics data, including mRNA, miRNA, DNA methylation, and proteomics data to discover altered biological functions in the samples being studied. Paintomics can integrate any type of functional genomics data into pathway analysis, provided that the features can be mapped onto genes or metabolites [162]. 3Omics [168] integrates transcriptomics, metabolomics and proteomics data into regulatory networks.

In all cases, integration of different datasets is rarely straightforward because each data type is analyzed separately with its own tailored algorithms that yield results in different formats. Tools that facilitate format conversions and the extraction of relevant results can help; examples of such workflow construction software packages include Anduril [169], Galaxy [170] and Chipster [171]. Anduril was developed for building complex pipelines with large datasets that require automated parallelization. The strength of Galaxy and Chipster is their usability; visualization is a key component of their design. Simultaneous or integrative visualization of the data in a genome browser is extremely useful for both data exploration and interpretation of results. Browsers can display in tandem mappings from most next-generation sequencing technologies, while adding custom tracks such as gene annotation, nucleotide variation or ENCODE datasets. For proteomics integration, the PG Nexus pipeline [172] converts mass spectrometry data to mappings that are co-visualized with RNA-seq alignments.

Outlook

RNA-seq has become the standard method for transcriptome analysis, but the technology and tools are continuing to evolve. It should be noted that the agreement between results obtained from different tools is still unsatisfactory and that results are affected by parameter settings, especially for genes that are expressed at low levels. The two major highlights in the current application of RNA-seq are the construction of transcriptomes from small amounts of starting materials and better transcript identification from longer reads. The state of the art in both of these areas is changing rapidly, but we will briefly outline what can be done now and what can be expected in the near future.

Single-cell RNA-seq

Single-cell RNA-seq (scRNA-seq) is one of the newest and most active fields of RNA-seq with its unique set of opportunities and challenges. Newer protocols such as Smart-seq [173] and Smart-seq2 [174] have enabled us to work from very small amounts of starting mRNA that, with proper amplification, can be obtained from just a single cell. The resulting single-cell libraries enable the identification of new, uncharacterized cell types in tissues. They also make it possible to measure a fascinating phenomenon in molecular biology, the stochasticity of gene expression in otherwise identical cells within a defined population. In this context, single cell studies are meaningful only when a set of individual cell libraries are compared with the cell population, with the aim of identifying subgroups of multiple cells with distinct combinations of expressed genes. Differences may be due to naturally occurring factors such as stage of the cell cycle, or may reflect rare cell types such as cancer stem cells. Recent rapid progress in methodologies for single-cell preparation, including the availability of single-cell platforms such as the Fluidigm C1 [8], has increased the number of individual cells analyzed from a handful to 50–90 per condition up to 800 cells at a time. Other methods, such as DROP-seq [175], can profile more than 10,000 cells at a time. This increased number of single-cell libraries in each experiment directly allows for the identification of smaller subgroups within the population.

The small amount of starting material and the PCR amplification limit the depth to which single-cell libraries can be sequenced productively, often to less than a million reads. Deeper sequencing for scRNA-seq will do little to improve quantification as the number of individual mRNA molecules in a cell is small (in the order of 100–300,000 transcripts) and only a fraction of them are successfully reverse-transcribed to cDNA [8, 176]; but deeper sequencing is potentially useful for discovering and measuring allele-specific expression, as additional reads could provide useful evidence.

Single-cell transcriptomes typically include about 3000–8000 expressed genes, which is far fewer than are counted in the transcriptomes of the corresponding pooled populations. The challenge is to distinguish the technical noise that results from a lack of sensitivity at the single-molecule level [173] (where capture rates of around 10–50 % result in the frequent loss of the most lowly expressed transcripts) from true biological noise where a transcript might not be transcribed and present in the cell for a certain amount of time while the protein is still present. The inclusion of added reference transcripts and the use of unique molecule identifiers (UMIs) have been applied to overcome amplification bias and to improve gene quantification [177, 178]. Methods that can quantify gene-level technical variation allow us to focus on biological variation that is likely to be of interest [179]. Typical quality-control steps involve setting aside libraries that contain few reads, libraries that have a low mapping rate, and libraries that have zero expression levels for housekeeping genes, such as *GAPDH* and *ACTB*, that are expected to be expressed at a detectable level.

Depending on the chosen single-cell protocol and the aims of the experiment, different bulk RNA-seq pipelines and tools can be used for different stages of the analysis as reviewed by Stegle et al. [180]. Single-cell libraries are typically analyzed by mapping to a reference transcriptome (using a program such as RSEM) without any attempt at new transcript discovery, although at least one package maps to the genome (Monocle [181]). While mapping onto the genome does result in a higher overall read-mapping rate, studies that are focused on gene expression alone with fewer reads per cell tend to use mapping to the reference transcriptome for the sake of simplicity. Other single-cell methods have been developed to measure single-cell DNA methylation [182] and single-cell open chromatin using ATAC-seq [183, 184]. At present, we can measure only one functional genomic data-type at a time in the same single cell, but we can expect that in the near future we will be able to recover the transcriptome of a single cell simultaneously with additional functional data.

Long-read sequencing

The major limitation of short-read RNA-seq is the difficulty in accurately reconstructing expressed full-length transcripts from the assembly of reads. This is particularly complicated in complex transcriptomes, where different but highly similar isoforms of the same gene are expressed, and for genes that have many exons and possible alternative promoters or 3' ends. Long-read technologies, such as Pacific-Biosciences (PacBio) SMRT and Oxford Nanopore, that were initially applied to genome sequencing are now being used for transcriptomics and have the potential to overcome this assembly problem. Long-read sequencing provides amplification-free, single-molecule sequencing of cDNAs that enables recovery of full-length transcripts without the need for an assembly step. PacBio adds adapters to the cDNA molecule and creates a circularized structure that can be sequenced with multiple passes within one single long read. The Nanopore GridION system can directly sequence RNA strands by using RNA processive enzymes and RNA-specific bases. Another interesting technology was previously known as Moleculo (now Illumina's TruSeq synthetic long-read technology), where Illumina library preparation is multiplexed and restricted to a limited number of long DNA molecules that are separately bar-coded and pooled back for sequencing. As one barcode corresponds to a limited number of molecules, assembly is greatly simplified and unambiguous reconstruction to long contigs is possible. This approach has recently been published for RNA-seq analysis [185].

PacBio RNA-seq is the long-read approach with the most publications to date. The technology has proven useful for unraveling isoform diversity at complex loci [186], and for determining allele-specific expression from single reads [187]. Nevertheless, long-read sequencing has its own set of limitations, such as a still high error rate that limits de novo transcript identifications and forces the technology to leverage the reference genome [188]. Moreover, the relatively low throughput of SMRT cells hampers the quantification of transcript expression. These two limitations can be addressed by matching PacBio experiments with regular, short-read RNA-seq. The accurate and abundant Illumina reads can be used both to correct long-read sequencing errors and to quantify transcript levels [189]. Updates in PacBio chemistry are increasing sequencing lengths to produce reads with a sufficient number of passes over the cDNA molecule to autocorrect sequencing errors. This will eventually improve sequencing accuracy and allow for genome-free determination of isoform-resolved transcriptomes.

Additional file

Additional file 1: Figure S1. Screenshots of RNA-seq data visualization. **a** Integrative Genomics Viewer (IGV) [77] display of a gene detected as differentially expressed between the two groups of samples by DEGseq [62]. The bottom track in the right panel is the gene annotation. The tracks are five samples from each group. **b** RNAseqViewer [80] display of the same data as in (**a**). **c** RNAseqViewer heatmap display of a gene detected as differentially spliced between two groups by both DSGSeq [70] and DEXSeq [69]. Introns are hidden in the display to emphasize the signals on the exons. **d** MISO [81] display of another gene detected as differentially spliced, with junction reads illustrated. (PDF 1152 kb)

Abbreviations

ASM: Alternative splicing module; ChIP-seq: Chromatin immunoprecipitation sequencing; DEG: Differentially expressed genes; eQTL: Expression quantitative loci; FDR: False discovery rate; FPKM: Fragments per kilobase of exon model per million mapped reads; GO: Gene Ontology; GSEA: Gene set enrichment analysis; GTF: Gene transfer format; IG: Immunoglobulin; IGV: Integrative Genomics Viewer; miRNA: MicroRNA; mRNA: Messenger RNA; PCA: Principal component analysis; PE read: Paired-end read; RNA-seq: RNA-sequencing; RPKM: Reads per kilobase of exon model per million reads; rRNA: Ribosomal RNA; RSEM: RNA-Seq by Expectation Maximization; scRNA-seq: Single-cell RNA-seq; SE read: Single-end read; siRNA: Short-interfering RNA; SNP: Single nucleotide polymorphism; sQTL: Splicing quantitative trait loci; sRNA: Small RNA; TF: Transcription factor; TPM: Transcripts per million.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ACo, PM and AM conceived the idea and shaped the structure of the manuscript. ACo drafted the experimental design, alignment and functional profiling sections and integrated contributions from all authors. PM drafted the visualization and de novo transcript reconstruction sections, and coordinated author contributions. ST drafted the quality-control and differential expression sections. DGC drafted the experimental design and integration sections. ACe contributed to drafting the integration section. AMP drafted the transcript fusion section. MWS drafted the small RNA section. DG drafted the eQTL section. LLE drafted the software comparison for differential expression section. LLE and XZ drafted the transcript isoform analysis sections. XZ contributed to drafting the visualization section. AM drafted the introduction and outlook sections and globally edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to thank Michael Love and Harold Pimentel for helpful suggestions on the initial draft of the manuscript. AC, ST, AM, DGC were supported by the FP7 STATegra project (grant 36000). Research in AC's laboratory was supported by MINECO grant BIO2012-40244 and co-funded with European Regional Development Funds (ERDF). Research in PM's laboratory is supported by ERC starting grant Relieve-IMDs and by a core support grant from the Wellcome Trust and MRC to the Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute. XZ was supported by the National Basic Research Program of China (2012CB316504). LLE was supported by JDRF (grant number 2-2013-32) and by the Sigrid Juselius Foundation. ACe was supported by the Academy of Finland (Center of Excellence in Cancer Genetics Research).

Author details
[1]Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32603, USA. [2]Centro de Investigación Príncipe Felipe, Genomics of Gene Expression Laboratory, 46012 Valencia, Spain. [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, Anne McLaren Laboratory for Regenerative Medicine, Department of Surgery, University of Cambridge, Cambridge CB2 0SZ, UK. [5]Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, 46020 Valencia, Spain. [6]Unit of Computational Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, 171 77 Stockholm, Sweden. [7]Center for Molecular Medicine, Karolinska Institutet, 17177 Stockholm, Sweden. [8]Unit of Clinical Epidemiology, Department of Medicine, Karolinska University Hospital, L8, 17176 Stockholm, Sweden. [9]Science for Life Laboratory, 17121 Solna, Sweden. [10]Systems Biology Laboratory, Institute of Biomedicine and Genome-Scale Biology Research Program, University of Helsinki, 00014 Helsinki, Finland. [11]School of Computing Science, Simon Fraser University, Burnaby V5A 1S6BC, Canada. [12]Department of Bioinformatics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University in Poznań, 61-614 Poznań, Poland. [13]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, FI-20520 Turku, Finland. [14]Key Lab of Bioinformatics/Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University, Beijing 100084, China. [15]School of Life Sciences, Tsinghua University, Beijing 100084, China. [16]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697-2300, USA. [17]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA 92697, USA.



References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:1–8.
2. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010;7:709–15.
3. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37:e123.
4. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7:1009–15.
5. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8:469–77.
6. abaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics. 2011;27:i383–91.
7. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
8. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32:1053–8.
9. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776–9.
10. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21:2213–23.
11. Andrews S. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 29 September 2014.
12. Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;11 Suppl 4:S7.
13. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 12 January 2016.
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
15. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
16. Picard. http://picard.sourceforge.net/. Accessed 12 January 2016.
17. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.
18. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28:2678–9.
19. Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43:e140.
20. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-seq data. BMC Bioinformatics. 2011;12:480.
21. Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.
22. Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ, et al. Genome-guided transcript assembly by integrative analysis of RNA sequence data. Nat Biotechnol. 2014;32:341–6.
23. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27:2325–9.
24. Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, Shah S, et al. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res. 2013;23:519–29.
25. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A. 2011;108:19867–72.
26. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
27. Hiller D, Wong WH. Simultaneous isoform discovery and quantification from RNA-Seq. Stat Biosci. 2013;5:100–18.
28. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.
29. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10:1185–91.
30. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660–6.
31. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.
32. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
33. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
34. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462–4.
35. Anders S, Pyl PT, Huber W. HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
36. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

37. UCSC Genome Bioinformatics: Frequently Asked Questions: Data File Formats. https://genome.ucsc.edu/FAQ/FAQformat.html#format4. Accessed on 12 January 2016.
38. Pachter L. Models for transcript quantification from RNA-seq. arXiv.org. 2011. http://arxiv.org/abs/1104.3889. Accessed 6 January 2016.
39. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
40. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
41. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013;10:71–3.
42. Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal RNA-Seq quantification with kallisto. https://liorpachter.wordpress.com/2015/05/10/near-optimal-rna-seq-quantification-with-kallisto/. Accessed 6 January 2016.
43. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12:R22.
44. Ma X, Zhang X. NURD: an implementation of a new method to estimate isoform expression from non-uniform RNA-seq data. BMC Bioinformatics. 2013;14:220.
45. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94.
46. Hansen K, Brenner S, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38:e131.
47. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
48. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
49. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13:523–38.
50. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2012;31:46–53.
51. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. Genetics. 2010;185:405–16.
52. Johnson WE, Rabinovic A, Li C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics. 2007;8:118–27.
53. Nueda MJ, Ferrer A, Conesa A. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. Biostatistics. 2012;13:553–66.
54. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23:2881–7.
55. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
56. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91.
57. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
58. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
59. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422.
60. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29:1035–43.
61. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22:519–36.
62. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010;26:136–8.
63. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14:R95.
64. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16:59–70.
65. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res. 2009;37:e75.
66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.
67. Singh D, Orellana CF, Hu Y, Jones CD, Liu Y, Chiang DY, et al. FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics. 2011;27:2633–40.
68. Shi Y, Jiang H. rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. PLoS One. 2013;8:e79448.
69. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012;22:2008–17.
70. Wang W, Qin Z, Feng Z, Wang X, Zhang X. Identifying differentially spliced genes from two groups of RNA-seq samples. Gene. 2013;518:164–70.
71. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111:E5593–601.
72. Drewe P, Stegle O, Hartmann L, Kahles A, Bohnert R, Wachter A, et al. Accurate detection of differential RNA processing. Nucleic Acids Res. 2013;41:5189–98.
73. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, et al. DissSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res. 2013;41:e39.
74. Hilker R, Stadermann KB, Doppmeier D, Kalinowski J, Stoye J, Straube J, et al. ReadXplorer - visualization and analysis of mapped sequences. Bioinformatics. 2014;30:2247–54.
75. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.
76. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high- performance genomics data visualization and exploration. Brief Bioinformatics. 2012;14:178–92.
77. Medina I, Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, et al. Genome Maps, a new generation genome browser. Nucleic Acids Res. 2013;41(Web Server issue):W41–6.
78. Fiume M, Williams V, Brook A, Brudno M. Savant: genome browser for high-throughput sequencing data. Bioinformatics. 2010;26:1938–44.
79. Rogé X, Zhang X. RNAseqViewer: visualization tool for RNA-Seq data. Bioinformatics. 2013;30:891–2.
80. Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdóttir H, et al. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics. 2015;31:2400–2.
81. Wu E, Nance T, Montgomery SB. SplicePlot: a utility for visualizing splicing quantitative trait loci. Bioinformatics. 2014;30:1025–6.
82. Ryan MC, Cleland J, Kim R, Wong WC, Weinstein JN. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics. 2012;28:2385–7.
83. Liu Q, Chen C, Shen E, Zhao F, Sun Z, Wu J. Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer. Genomics. 2012;99:178–82.
84. Dietrich S, Wiegand S, Liesegang H. TraV: a genome context sensitive transcriptome browser. PLoS One. 2014;9:e93677.
85. Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, et al. State-of-the-art fusion-finder algorithms sensitivity and specificity. BioMed Res Int. 2013;15:340620.
86. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A. 2009;106:12353–8.
87. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene. 2015;34:4845–54.
88. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-seq data. PLoS Comput Biol. 2011;7:e1001138.
89. Wu C, Wyatt AW, McPherson A, Lin D, McConeghy BJ, Mo F, et al. Poly-gene fusion transcripts and chromothripsis in prostate cancer. Gene Chromosomes Cancer. 2012;51:1144–53.

90. Wyatt AW, Mo F, Wang K, McConeghy B, Brahmbhatt S, Jong L, et al. Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. Genome Biol. 2014;15:426.
91. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nat Commun. 2014;5:4846.
92. Rabbitts TH. Commonality but diversity in cancer gene fusions. Cell. 2009;137:391–5.
93. McPherson A, Wu C, Hajirasouliha I, Hormozdiari F, Hach F, Lapuk A, et al. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics. 2011;27:1481–8.
94. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27:2903–4.
95. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 2011;21:56–67.
96. Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, et al. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. RNA Biol. 2013;10:1087–92.
97. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
98. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
99. Prüfer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J. PatMaN: rapid alignment of short sequences to large databases. Bioinformatics. 2008;24:1530–1.
100. Emde AK, Grunert M, Weese D, Reinert K, Sperling SR. MicroRazerS: rapid alignment of small RNA reads. Bioinformatics. 2010;26:123–4.
101. An J, Lai J, Lehman ML, Nelson CC. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res. 2013;41:727–37.
102. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics. 2011;27:2614–5.
103. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, et al. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics. 2012;28:2059–61.
104. Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. RNA. 2013;19:740–51.
105. Giurato G, De Filippo MR, Rinaldi A, Hashim A, Nassa G, Ravo M, et al. iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. BMC Bioinformatics. 2013;14:362.
106. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11:1–12.
107. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14:7.
108. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. BMC Bioinformatics. 2013;14 Suppl 5:S16.
109. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
110. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.
111. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat Protocols. 2009;4:44–57.
112. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
113. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, et al. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010;38 suppl 2:W210–3.
114. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: juggling between evolution and stability. Brief Bioinformatics. 2004;5:39–55.
115. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. The Pfam protein families database. Nucleic Acids Res. 2014;42(Database issue):D222–30.
116. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2011;40(Database issue):D306–12.
117. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
118. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res. 2009;37 suppl 1:D136–40.
119. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42(Database issue):D68–73.
120. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in *Drosophila*. Genome Biol. 2003;5:R1.
121. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383.
122. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007;448:470–3.
123. Gilad Y, Rifkin S, Pritchard J. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008;24:408–15.
124. Gaffney D. Global properties and functional complexity of human gene regulatory variation. PLoS Genet. 2013;9:e1003501.
125. Montgomery S, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
126. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
127. Lappalainen T, Sammeth M, Friedlander M, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
128. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, Shi J, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24:14–24.
129. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010;11:533–8.
130. Sun W. A statistical framework for eQTL mapping using RNA-seq data. Biometrics. 2012;68:1–11.
131. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12:1061–3.
132. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat Genet. 2015. doi: 10.1038/ng.3467.
133. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. Proc Natl Acad Sci U S A. 2003;100:9440–5.
134. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005;1:e78.
135. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014;344:519–23.
136. Shabalin A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
137. Louhimo R, Lepikhova T, Monni O, Hautaniemi S. Comparative analysis of algorithms for integration of copy number and expression data. Nat Methods. 2012;9:351–5.
138. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. 2011;21:1028–41.
139. Li JL, Mazar J, Zhong C, Faulkner GJ, Govindarajan SS, Zhang Z, et al. Genome-wide methylated CpG island profiles of melanoma cells reveal a melanoma coregulation network. Sci Rep. 2013;3:2962.
140. Xie L, Weichel B, Ohm JE, Zhang K. An integrative analysis of DNA methylation and RNA-Seq data for human heart, kidney and liver. BMC Syst Biol. 2011;5 Suppl 3:S4.
141. Van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics. 2012;13:636.

142. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.
143. Yeang C-H. An integrated analysis of molecular aberrations in NCI-60 cell lines. BMC Bioinformatics. 2010;11:495.
144. Jeong J, Li L, Liu Y, Nephew KP, Huang YHM, Shen C. An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer. BMC Med Genomics. 2010;3:55.
145. Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. Bioinformatics. 2014;30:2360–6.
146. Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc. 2013;8:2502–15.
147. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
148. Madrigal P, Krajewski P. Uncovering correlated variability in epigenomic datasets using the Karhunen-Loeve transform. BioData Min. 2015;8:20.
149. Angelini C, Costa V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. Front Cell Dev Biol. 2014;2:51.
150. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150:1274–86.
151. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk - database: prediction of possible miRNA binding sites by 'walking' the genes of 3 genomes. J Biomed Inform. 2011;44:839–47.
152. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008;36:D154–8.
153. Wu X, Watson M. CORNA: testing gene lists for regulation by microRNAs. Bioinformatics. 2009;25:832–3.
154. Lee H, Yang Y, Chae H, Nam S, Choi D, Tangchaisin P, et al. BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2. IEEE Trans Nanobiosci. 2012;11:266–72.
155. Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. Nucleic Acids Res. 2009;37:W356–62.
156. Sales G, Coppe A, Bisognin A, Bortoluzzi S, Romualdi C. MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. Nucleic Acids Res. 2010;38:W352–9.
157. Icay K, Chen P, Cervera C, Lehtonen R, Hautaniemi S. SePIA: RNA and smallRNA-sequence processing, integration, and analysis. 2015. http://anduril.org/sepia. Accessed 6 Jan 2016.
158. de Sousa AR, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009;5:1512–26.
159. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13:227–32.
160. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 2013;5:1469–78.
161. Suhre K, Schmitt-Kopplin P. MassTRIX: mass translator into pathways. Nucleic Acids Res. 2008;36(Web Server issue):W481–4.
162. García-Alcalde F, García-López F, Dopazo J, Conesa A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics. 2011;27:137–9.
163. Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, Klapperstück M, et al. VANTED v2: a framework for systems biology applications. BMC Syst Biol. 2012;6:139.
164. Tuncbag N, McCallum S, Huang SS, Fraenkel E. SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. Nucleic Acids Res. 2012;40:W505–9.
165. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011;27:i401–9.
166. Le H-S, Bar-Joseph Z. Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation. Bioinformatics. 2013;29:i89–97.
167. Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. Nucleic Acids Res. 2013;41:9622–33.
168. Kuo TC, Tian TF, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013;7:64.
169. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med. 2010;2:65.
170. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:R86.
171. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12:507.
172. Pang CNI, Tay AP, Aya C. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. J Proteome Res. 2014;13:84–98.
173. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82.
174. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
175. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
176. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24:496–510.
177. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11:163–6.
178. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011;9:72–4.
179. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10:1093–5.
180. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16:133–45.
181. Trapnell C, Cacchiarelli D. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
182. Lorthongpanich C, Cheow LF, Balu S, Quake SR, Knowles BB, Burkholder WF, et al. Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. Science. 2013;341:1110–2.
183. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90.
184. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348:910–4.
185. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015;33:736–42.
186. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci U S A. 2013;110:E4821–30.
187. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci U S A. 2014;111:9869–74.
188. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. PLoS One. 2012;7:e46679.
189. Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. Nat Biotechnol. 2011;29:572–3.
190. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. J Comput Biol. 2013;20:970–8.
191. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics. 2013;29:656–7.

192. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4:14.
193. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009;27:1135–7.
194. Robinson DG, Storey JD. subSeq: determining appropriate sequencing depth through efficient read subsampling. Bioinformatics. 2014;30:3424–6.
195. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2013;30:301–4.
196. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014;32:903–14.
197. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21:1543–51.
198. Kouzine F, Wojtowicz D, Yamane A, Resch W, Kieffer-Kwon KR, Bandle R, et al. Global regulation of promoter melting in naive lymphocytes. Cell. 2013;153:988–99.
199. Van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res. 2014;322:12–20.
200. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
201. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
202. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–81.
203. Jean G, Kahles A, Sreedharan VT, De Bona F, Rätsch G. RNA-Seq read alignments with PALMapper. Curr Protoc Bioinformatics. 2010;11(6).
204. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38:e178.
205. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012;9:1185–8.
206. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97.
207. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
208. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot. 2012;99:248–56.
209. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics. 2012;13:484.
210. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. Nucleic Acids Res. 2012;40:10084–97.
211. Seyednasrollah F, Rantanen K, Jaakkola P, Elo LL. ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. 2016;44(1):e1. doi:10.1093/nar/gkv806.
212. Bi Y, Davuluri RV. NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:262.
213. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics. 2014;30:2598–602.
214. GTEx Consortium. The Genotype-Tissue expression (GTEx) project. Nat Genet. 2013;45:580–5.

Love *et al. Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8

Genome Biology

METHOD Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

Abstract

In comparative high-throughput sequencing assays, a fundamental task is the analysis of count data, such as read counts per gene in RNA-seq, for evidence of systematic changes across experimental conditions. Small replicate numbers, discreteness, large dynamic range and the presence of outliers require a suitable statistical approach. We present *DESeq2*, a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. This enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression. The *DESeq2* package is available at http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html.

Background

The rapid adoption of high-throughput sequencing (HTS) technologies for genomic studies has resulted in a need for statistical methods to assess quantitative differences between experiments. An important task here is the analysis of RNA sequencing (RNA-seq) data with the aim of finding genes that are differentially expressed across groups of samples. This task is general: methods for it are typically also applicable for other comparative HTS assays, including chromatin immunoprecipitation sequencing, chromosome conformation capture, or counting observed taxa in metagenomic studies.

Besides the need to account for the specifics of count data, such as non-normality and a dependence of the variance on the mean, a core challenge is the small number of samples in typical HTS experiments – often as few as two or three replicates per condition. Inferential methods that treat each gene separately suffer here from lack of power, due to the high uncertainty of within-group variance estimates. In high-throughput assays, this limitation can be overcome by pooling information across genes, specifically, by exploiting assumptions about the similarity of the variances of different genes measured in the same experiment [1].

*Correspondence: sanders@fs.tum.de
[2]Genome Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
Full list of author information is available at the end of the article

Many methods for differential expression analysis of RNA-seq data perform such information sharing across genes for variance (or, equivalently, dispersion) estimation. *edgeR* [2,3] moderates the dispersion estimate for each gene toward a common estimate across all genes, or toward a local estimate from genes with similar expression strength, using a weighted conditional likelihood. Our *DESeq* method [4] detects and corrects dispersion estimates that are too low through modeling of the dependence of the dispersion on the average expression strength over all samples. *BBSeq* [5] models the dispersion on the mean, with the mean absolute deviation of dispersion estimates used to reduce the influence of outliers. *DSS* [6] uses a Bayesian approach to provide an estimate for the dispersion for individual genes that accounts for the heterogeneity of dispersion values for different genes. *baySeq* [7] and *ShrinkBayes* [8] estimate priors for a Bayesian model over all genes, and then provide posterior probabilities or false discovery rates (FDRs) for differential expression.

The most common approach in the comparative analysis of transcriptomics data is to test the null hypothesis that the logarithmic fold change (LFC) between treatment and control for a gene's expression is exactly zero, i.e., that the gene is not at all affected by the treatment. Often the goal of differential analysis is to produce a list of genes passing multiple-test adjustment, ranked by *P* value. However, small changes, even if statistically highly significant, might not be the most interesting candidates for

BioMed Central

© 2014 Love et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

further investigation. Ranking by fold change, on the other hand, is complicated by the noisiness of LFC estimates for genes with low counts. Furthermore, the number of genes called significantly differentially expressed depends as much on the sample size and other aspects of experimental design as it does on the biology of the experiment – and well-powered experiments often generate an overwhelmingly long list of hits [9]. We, therefore, developed a statistical framework to facilitate gene ranking and visualization based on stable estimation of effect sizes (LFCs), as well as testing of differential expression with respect to user-defined thresholds of biological significance.

Here we present *DESeq2*, a successor to our *DESeq* method [4]. *DESeq2* integrates methodological advances with several novel features to facilitate a more quantitative analysis of comparative RNA-seq data using shrinkage estimators for dispersion and fold change. We demonstrate the advantages of *DESeq2*'s new features by describing a number of applications possible with shrunken fold changes and their estimates of standard error, including improved gene ranking and visualization, hypothesis tests above and below a threshold, and the regularized logarithm transformation for quality assessment and clustering of overdispersed count data. We furthermore compare *DESeq2*'s statistical power with existing tools, revealing that our methodology has high sensitivity and precision, while controlling the false positive rate. *DESeq2* is available [10] as an R/Bioconductor package [11].

Results and discussion

Model and normalization

The starting point of a *DESeq2* analysis is a count matrix K with one row for each gene i and one column for each sample j. The matrix entries K_{ij} indicate the number of sequencing reads that have been unambiguously mapped to a gene in a sample. Note that although we refer in this paper to counts of reads in genes, the methods presented here can be applied as well to other kinds of HTS count data. For each gene, we fit a generalized linear model (GLM) [12] as follows.

We model read counts K_{ij} as following a negative binomial distribution (sometimes also called a gamma-Poisson distribution) with mean μ_{ij} and dispersion α_i. The mean is taken as a quantity q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, scaled by a normalization factor s_{ij}, i.e., $\mu_{ij} = s_{ij}q_{ij}$. For many applications, the same constant s_j can be used for all genes in a sample, which then accounts for differences in sequencing depth between samples. To estimate these *size factors*, the *DESeq2* package offers the median-of-ratios method already used in *DESeq* [4]. However, it can be advantageous to calculate gene-specific normalization factors s_{ij} to account for further sources of technical biases such as differing dependence on GC content, gene length or the like, using published methods [13,14], and these can be supplied instead.

We use GLMs with a logarithmic link, $\log_2 q_{ij} = \sum_r x_{jr}\beta_{ir}$, with design matrix elements x_{jr} and coefficients β_{ir}. In the simplest case of a comparison between two groups, such as treated and control samples, the design matrix elements indicate whether a sample j is treated or not, and the GLM fit returns coefficients indicating the overall expression strength of the gene and the $\log_2$ fold change between treatment and control. The use of linear models, however, provides the flexibility to also analyze more complex designs, as is often useful in genomic studies [15].

Empirical Bayes shrinkage for dispersion estimation

Within-group variability, i.e., the variability between replicates, is modeled by the dispersion parameter α_i, which describes the variance of counts via $\mathrm{Var}\, K_{ij} = \mu_{ij} + \alpha_i\mu_{ij}^2$. Accurate estimation of the dispersion parameter α_i is critical for the statistical inference of differential expression. For studies with large sample sizes this is usually not a problem. For controlled experiments, however, sample sizes tend to be smaller (experimental designs with as little as two or three replicates are common and reasonable), resulting in highly variable dispersion estimates for each gene. If used directly, these noisy estimates would compromise the accuracy of differential expression testing.

One sensible solution is to share information across genes. In *DESeq2*, we assume that genes of similar average expression strength have similar dispersion. We here explain the concepts of our approach using as examples a dataset by Bottomly *et al.* [16] with RNA-seq data for mice of two different strains and a dataset by Pickrell *et al.* [17] with RNA-seq data for human lymphoblastoid cell lines. For the mathematical details, see Methods.

We first treat each gene separately and estimate gene-wise dispersion estimates (using maximum likelihood), which rely only on the data of each individual gene (black dots in Figure 1). Next, we determine the location parameter of the distribution of these estimates; to allow for dependence on average expression strength, we fit a smooth curve, as shown by the red line in Figure 1. This provides an accurate estimate for the expected dispersion value for genes of a given expression strength but does not represent deviations of individual genes from this overall trend. We then shrink the gene-wise dispersion estimates toward the values predicted by the curve to obtain final dispersion values (blue arrow heads). We use an empirical Bayes approach (Methods), which lets the strength of shrinkage depend (i) on an estimate of how close true dispersion values tend to be to the fit and (ii) on the degrees of freedom: as the sample size increases, the shrinkage decreases in strength, and eventually becomes negligible. Our approach therefore accounts for gene-specific

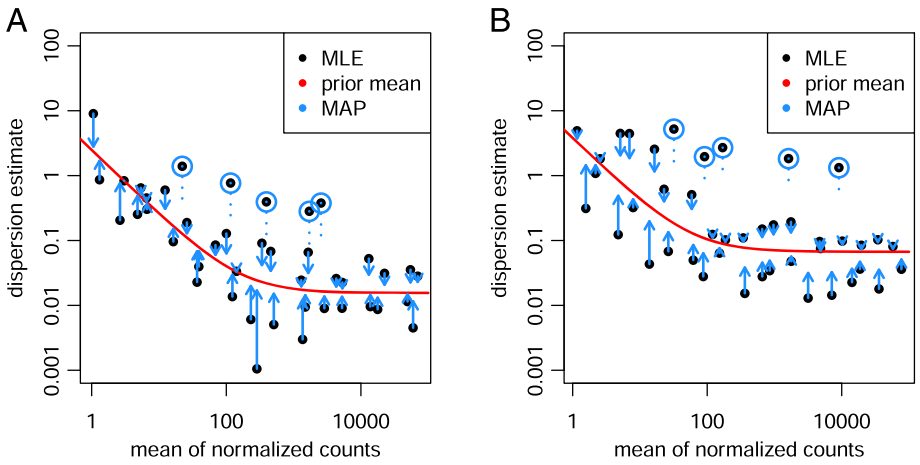


Figure 1 Shrinkage estimation of dispersion. Plot of dispersion estimates over the average expression strength **(A)** for the Bottomly *et al.* [16] dataset with six samples across two groups and **(B)** for five samples from the Pickrell *et al.* [17] dataset, fitting only an intercept term. First, gene-wise MLEs are obtained using only the respective gene's data (black dots). Then, a curve (red) is fit to the MLEs to capture the overall trend of dispersion-mean dependence. This fit is used as a prior mean for a second estimation round, which results in the final MAP estimates of dispersion (arrow heads). This can be understood as a shrinkage (along the blue arrows) of the noisy gene-wise estimates toward the consensus represented by the red line. The black points circled in blue are detected as dispersion outliers and not shrunk toward the prior (shrinkage would follow the dotted line). For clarity, only a subset of genes is shown, which is enriched for dispersion outliers. Additional file 1: Figure S1 displays the same data but with dispersions of all genes shown. MAP, maximum *a posteriori*; MLE, maximum-likelihood estimate.

variation to the extent that the data provide this information, while the fitted curve aids estimation and testing in less information-rich settings.

Our approach is similar to the one used by *DSS* [6], in that both methods sequentially estimate a prior distribution for the true dispersion values around the fit, and then provide the maximum *a posteriori* (MAP) as the final estimate. It differs from the previous implementation of *DESeq*, which used the maximum of the fitted curve and the gene-wise dispersion estimate as the final estimate and tended to overestimate the dispersions (Additional file 1: Figure S2). The approach of *DESeq2* differs from that of *edgeR* [3], as *DESeq2* estimates the width of the prior distribution from the data and therefore automatically controls the amount of shrinkage based on the observed properties of the data. In contrast, the default steps in *edgeR* require a user-adjustable parameter, the *prior degrees of freedom*, which weighs the contribution of the individual gene estimate and *edgeR*'s dispersion fit.

Note that in Figure 1 a number of genes with gene-wise dispersion estimates below the curve have their final estimates raised substantially. The shrinkage procedure thereby helps avoid potential false positives, which can result from underestimates of dispersion. If, on the other hand, an individual gene's dispersion is far above the distribution of the gene-wise dispersion estimates of other genes, then the shrinkage would lead to a greatly reduced final estimate of dispersion. We reasoned that in many cases, the reason for extraordinarily high dispersion of a gene is that it does not obey our modeling assumptions; some genes may show much higher variability than others for biological or technical reasons, even though they have the same average expression levels. In these cases, inference based on the shrunken dispersion estimates could lead to undesirable false positive calls. *DESeq2* handles these cases by using the gene-wise estimate instead of the shrunken estimate when the former is more than 2 residual standard deviations above the curve.

Empirical Bayes shrinkage for fold-change estimation

A common difficulty in the analysis of HTS data is the strong variance of LFC estimates for genes with low read count. We demonstrate this issue using the dataset by Bottomly *et al.* [16]. As visualized in Figure 2A, weakly expressed genes seem to show much stronger differences between the compared mouse strains than strongly expressed genes. This phenomenon, seen in most HTS datasets, is a direct consequence of dealing with *count* data, in which ratios are inherently noisier when counts are low. This heteroskedasticity (variance of LFCs depending on mean count) complicates downstream analysis and data interpretation, as it makes effect sizes difficult to compare across the dynamic range of the data.

DESeq2 overcomes this issue by shrinking LFC estimates toward zero in a manner such that shrinkage is stronger when the available information for a gene is low, which may be because counts are low, dispersion is high or there are few degrees of freedom. We again employ an empirical Bayes procedure: we first perform

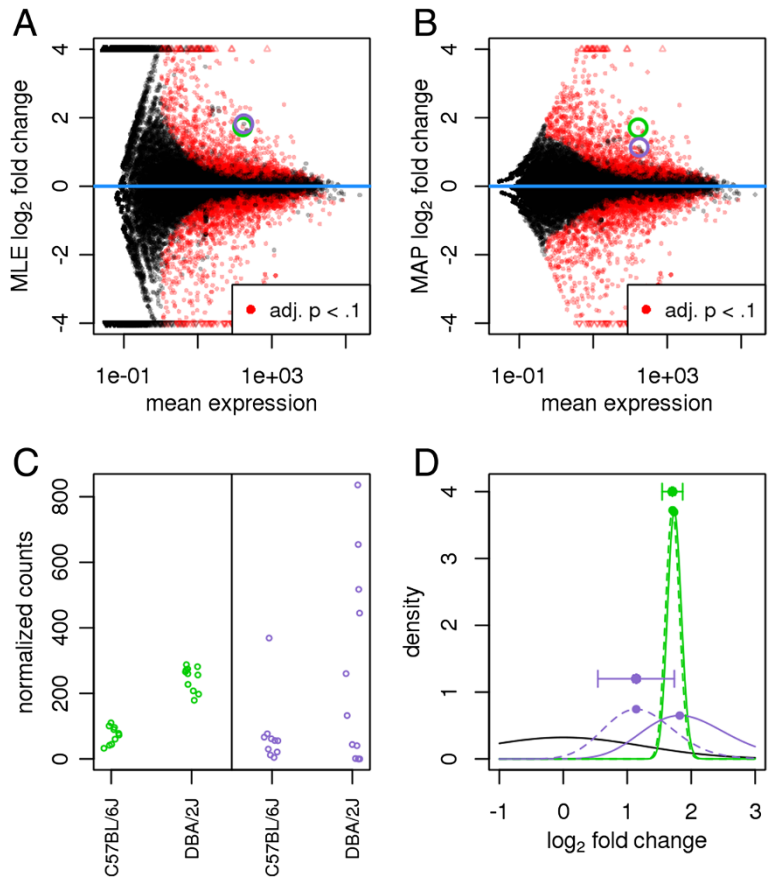


Figure 2 Effect of shrinkage on logarithmic fold change estimates. Plots of the **(A)** MLE (i.e., no shrinkage) and **(B)** MAP estimate (i.e., with shrinkage) for the LFCs attributable to mouse strain, over the average expression strength for a ten vs eleven sample comparison of the Bottomly *et al.* [16] dataset. Small triangles at the top and bottom of the plots indicate points that would fall outside of the plotting window. Two genes with similar mean count and MLE logarithmic fold change are highlighted with green and purple circles. **(C)** The counts (normalized by size factors s_j) for these genes reveal low dispersion for the gene in green and high dispersion for the gene in purple. **(D)** Density plots of the likelihoods (solid lines, scaled to integrate to 1) and the posteriors (dashed lines) for the green and purple genes and of the prior (solid black line): due to the higher dispersion of the purple gene, its likelihood is wider and less peaked (indicating less information), and the prior has more influence on its posterior than for the green gene. The stronger curvature of the green posterior at its maximum translates to a smaller reported standard error for the MAP LFC estimate (horizontal error bar). adj., adjusted; LFC, logarithmic fold change; MAP, maximum *a posteriori*; MLE, maximum-likelihood estimate.

ordinary GLM fits to obtain maximum-likelihood estimates (MLEs) for the LFCs and then fit a zero-centered normal distribution to the observed distribution of MLEs over all genes. This distribution is used as a prior on LFCs in a second round of GLM fits, and the MAP estimates are kept as final estimates of LFC. Furthermore, a standard error for each estimate is reported, which is derived from the posterior's curvature at its maximum (see Methods for details). These shrunken LFCs and their standard errors are used in the Wald tests for differential expression described in the next section.

The resulting MAP LFCs are biased toward zero in a manner that removes the problem of exaggerated LFCs for low counts. As Figure 2B shows, the strongest LFCs are no longer exhibited by genes with weakest expression. Rather, the estimates are more evenly spread around zero, and for very weakly expressed genes (with less than one read per sample on average), LFCs hardly deviate from zero, reflecting that accurate LFC estimates are not possible here.

The strength of shrinkage does not depend simply on the mean count, but rather on the amount of information available for the fold change estimation (as indicated by the observed Fisher information; see Methods). Two genes with equal expression strength but different dispersions will experience a different amount of shrinkage (Figure 2C,D). The shrinkage of LFC estimates can be described as a bias-variance trade-off [18]: for genes with little information for LFC estimation, a reduction of the strong variance is bought at the cost of accepting a bias toward zero, and this can result in an overall reduction in mean squared error, e.g., when comparing to LFC

estimates from a new dataset. Genes with high information for LFC estimation will have, in our approach, LFCs with both low bias and low variance. Furthermore, as the degrees of freedom increase, and the experiment provides more information for LFC estimation, the shrunken estimates will converge to the unshrunken estimates. We note that other Bayesian efforts toward moderating fold changes for RNA-seq include hierarchical models [8,19] and the *GFOLD* (or generalized fold change) tool [20], which uses a posterior distribution of LFCs.

The shrunken MAP LFCs offer a more reproducible quantification of transcriptional differences than standard MLE LFCs. To demonstrate this, we split the Bottomly *et al.* samples equally into two groups, I and II, such that each group contained a balanced split of the strains, simulating a scenario where an experiment (samples in group I) is performed, analyzed and reported, and then independently replicated (samples in group II). Within each group, we estimated LFCs between the strains and compared between groups I and II, using the MLE LFCs (Figure 3A) and using the MAP LFCs (Figure 3B). Because the shrinkage moves large LFCs that are not well supported by the data toward zero, the agreement between the two independent sample groups increases considerably. Therefore, shrunken fold-change estimates offer a more reliable basis for quantitative conclusions than normal MLEs.

This makes shrunken LFCs also suitable for ranking genes, e.g., to prioritize them for follow-up experiments. For example, if we sort the genes in the two sample groups of Figure 3 by unshrunken LFC estimates, and consider the 100 genes with the strongest up- or down-regulation in group I, we find only 21 of these again among the top 100 up- or down-regulated genes in group II. However, if we rank the genes by shrunken LFC estimates, the overlap improves to 81 of 100 genes (Additional file 1: Figure S3).

A simpler often used method is to add a fixed number (pseudocount) to all counts before forming ratios. However, this requires the choice of a tuning parameter and only reacts to one of the sources of uncertainty, low counts, but not to gene-specific dispersion differences or sample size. We demonstrate this in the *Benchmarks* section below.

Hypothesis tests for differential expression

After GLMs are fit for each gene, one may test whether each model coefficient differs significantly from zero. *DESeq2* reports the standard error for each shrunken LFC estimate, obtained from the curvature of the coefficient's posterior (dashed lines in Figure 2D) at its maximum. For significance testing, *DESeq2* uses a Wald test: the shrunken estimate of LFC is divided by its standard error, resulting in a z-statistic, which is compared to a standard normal distribution. (See Methods for details.) The Wald test allows testing of individual coefficients, or contrasts of coefficients, without the need to fit a reduced model as with the likelihood ratio test, though the likelihood ratio test is also available as an option in *DESeq2*. The Wald test P values from the subset of genes that pass an independent filtering step, described in the next section, are adjusted for multiple testing using the procedure of Benjamini and Hochberg [21].

Automatic independent filtering

Due to the large number of tests performed in the analysis of RNA-seq and other genome-wide experiments, the multiple testing problem needs to be addressed. A popular objective is control or estimation of the FDR. Multiple

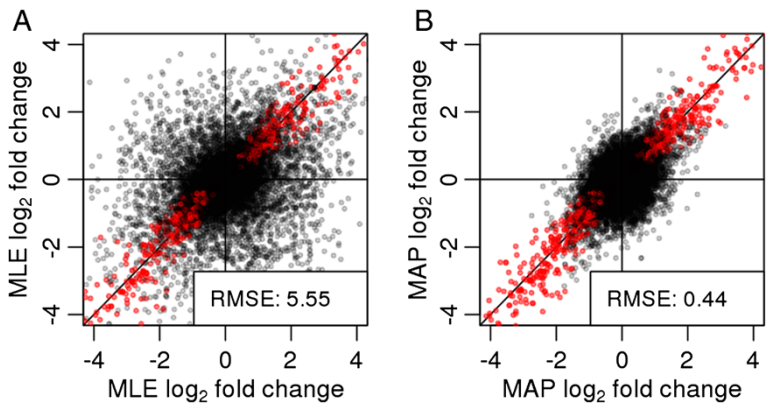


Figure 3 Stability of logarithmic fold changes. *DESeq2* is run on equally split halves of the data of Bottomly *et al.* [16], and the LFCs from the halves are plotted against each other. **(A)** MLEs, i.e., without LFC shrinkage. **(B)** MAP estimates, i.e., with shrinkage. Points in the top left and bottom right quadrants indicate genes with a change of sign of LFC. Red points indicate genes with adjusted P value < 0.1. The legend displays the root-mean-square error of the estimates in group I compared to those in group II. LFC, logarithmic fold change; MAP, maximum *a posteriori*; MLE, maximum-likelihood estimate; RMSE, root-mean-square error.

testing adjustment tends to be associated with a loss of power, in the sense that the FDR for a set of genes is often higher than the individual *P* values of these genes. However, the loss can be reduced if genes that have little or no chance of being detected as differentially expressed are omitted from the testing, provided that the criterion for omission is independent of the test statistic under the null hypothesis [22] (see Methods). *DESeq2* uses the average expression strength of each gene, across all samples, as its filter criterion, and it omits all genes with mean normalized counts below a filtering threshold from multiple testing adjustment. *DESeq2* by default will choose a threshold that maximizes the number of genes found at a user-specified target FDR. In Figures 2A,B and 3, genes found in this way to be significant at an estimated FDR of 10% are depicted in red. Depending on the distribution of the mean normalized counts, the resulting increase in power can be substantial, sometimes making the difference in whether or not any differentially expressed genes are detected.

Hypothesis tests with thresholds on effect size

Specifying minimum effect size

Most approaches to testing for differential expression, including the default approach of *DESeq2*, test against the null hypothesis of *zero* LFC. However, if any biological processes are genuinely affected by the difference in experimental treatment, this null hypothesis implies that the gene under consideration is *perfectly* decoupled from these processes. Due to the high interconnectedness of cells' regulatory networks, this hypothesis is, in fact, implausible, and arguably wrong for many if not most genes. Consequently, with sufficient sample size, even genes with a very small but non-zero LFC will eventually be detected as differentially expressed. A change should therefore be of sufficient magnitude to be considered *biologically significant*. For small-scale experiments, statistical significance is often a much stricter requirement than biological significance, thereby relieving the researcher from the need to decide on a threshold for biological significance.

For well-powered experiments, however, a statistical test against the conventional null hypothesis of zero LFC may report genes with statistically significant changes that are so weak in effect strength that they could be considered irrelevant or distracting. A common procedure is to disregard genes whose estimated LFC β_{ir} is below some threshold, $|\beta_{ir}| \leq \theta$. However, this approach loses the benefit of an easily interpretable FDR, as the reported *P* value and adjusted *P* value still correspond to the test of *zero* LFC. It is therefore desirable to include the threshold in the statistical testing procedure directly, i.e., not to filter post hoc on a reported fold-change *estimate*, but rather to evaluate statistically directly whether there is sufficient evidence that the LFC is above the chosen threshold.

DESeq2 offers tests for composite null hypotheses of the form $|\beta_{ir}| \leq \theta$, where β_{ir} is the shrunken LFC from the estimation procedure described above. (See Methods for details.) Figure 4A demonstrates how such a thresholded test gives rise to a curved decision boundary: to reach significance, the estimated LFC has to exceed the specified threshold by an amount that depends on the available information. We note that related approaches to generate gene lists that satisfy both statistical and biological significance criteria have been previously discussed for microarray data [23] and recently for sequencing data [19].

Specifying maximum effect size

Sometimes, a researcher is interested in finding genes that are not, or only very weakly, affected by the treatment or experimental condition. This amounts to a setting similar to the one just discussed, but the roles of the null and alternative hypotheses are swapped. We are here asking for evidence of the effect being weak, not for evidence of the effect being zero, because the latter question is rarely tractable. The meaning of *weak* needs to be quantified for the biological question at hand by choosing a suitable threshold θ for the LFC. For such analyses, *DESeq2* offers a test of the composite null hypothesis $|\beta_{ir}| \geq \theta$, which will report genes as significant for which there is evidence that their LFC is weaker than θ. Figure 4B shows the outcome of such a test. For genes with very low read count, even an estimate of zero LFC is not significant, as the large uncertainty of the estimate does not allow us to exclude that the gene may in truth be more than weakly affected by the experimental condition. Note the lack of LFC shrinkage: to find genes with weak differential expression, *DESeq2* requires that the LFC shrinkage has been disabled. This is because the zero-centered prior used for LFC shrinkage embodies a *prior* belief that LFCs tend to be small, and hence is inappropriate here.

Detection of count outliers

Parametric methods for detecting differential expression can have gene-wise estimates of LFC overly influenced by individual outliers that do not fit the distributional assumptions of the model [24]. An example of such an outlier would be a gene with single-digit counts for all samples, except one sample with a count in the thousands. As the aim of differential expression analysis is typically to find *consistently* up- or down-regulated genes, it is useful to consider diagnostics for detecting individual observations that overly influence the LFC estimate and *P* value for a gene. A standard outlier diagnostic is Cook's distance [25], which is defined within each gene for each sample as the scaled distance that the coefficient vector,

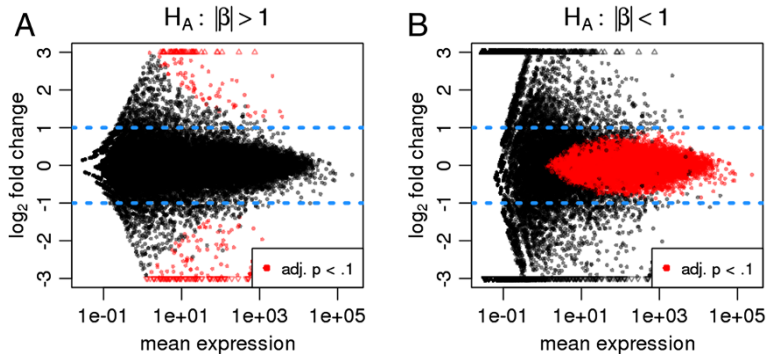


Figure 4 Hypothesis testing involving non-zero thresholds. Shown are plots of the estimated fold change over average expression strength ("minus over average", or MA-plots) for a ten vs eleven comparison using the Bottomly *et al.* [16] dataset, with highlighted points indicating low adjusted *P* values. The alternate hypotheses are that logarithmic (base 2) fold changes are **(A)** greater than 1 in absolute value or **(B)** less than 1 in absolute value. adj., adjusted.

$\vec{\beta}_i$, of a linear model or GLM would move if the sample were removed and the model refit.

DESeq2 flags, for each gene, those samples that have a Cook's distance greater than the 0.99 quantile of the $F(p, m-p)$ distribution, where p is the number of model parameters including the intercept, and m is the number of samples. The use of the F distribution is motivated by the heuristic reasoning that removing a single sample should not move the vector $\vec{\beta}_i$ outside of a 99% confidence region around $\vec{\beta}_i$ fit using all the samples [25]. However, if there are two or fewer replicates for a condition, these samples do not contribute to outlier detection, as there are insufficient replicates to determine outlier status.

How should one deal with flagged outliers? In an experiment with many replicates, discarding the outlier and proceeding with the remaining data might make best use of the available data. In a small experiment with few samples, however, the presence of an outlier can impair inference regarding the affected gene, and merely ignoring the outlier may even be considered data cherry-picking – and therefore, it is more prudent to exclude the whole gene from downstream analysis.

Hence, *DESeq2* offers two possible responses to flagged outliers. By default, outliers in conditions with six or fewer replicates cause the whole gene to be flagged and removed from subsequent analysis, including *P* value adjustment for multiple testing. For conditions that contain seven or more replicates, *DESeq2* replaces the outlier counts with an imputed value, namely the trimmed mean over all samples, scaled by the size factor, and then re-estimates the dispersion, LFCs and *P* values for these genes. As the outlier is replaced with the value predicted by the null hypothesis of no differential expression, this is a more conservative choice than simply omitting the outlier. When there are many degrees of freedom, the second approach avoids discarding genes that might contain true differential expression.

Additional file 1: Figure S4 displays the outlier replacement procedure for a single gene in a seven by seven comparison of the Bottomly *et al.* [16] dataset. While the original fitted means are heavily influenced by a single sample with a large count, the corrected LFCs provide a better fit to the majority of the samples.

Regularized logarithm transformation

For certain analyses, it is useful to transform data to render them homoskedastic. As an example, consider the task of assessing sample similarities in an unsupervised manner using a clustering or ordination algorithm. For RNA-seq data, the problem of heteroskedasticity arises: if the data are given to such an algorithm on the original count scale, the result will be dominated by highly expressed, highly variable genes; if logarithm-transformed data are used, undue weight will be given to weakly expressed genes, which show exaggerated LFCs, as discussed above. Therefore, we use the shrinkage approach of *DESeq2* to implement a *regularized logarithm* transformation (rlog), which behaves similarly to a $\log_2$ transformation for genes with high counts, while shrinking together the values for different samples for genes with low counts. It therefore avoids a commonly observed property of the standard logarithm transformation, the spreading apart of data for genes with low counts, where random noise is likely to dominate any biologically meaningful signal. When we consider the variance of each gene, computed across samples, these variances are stabilized – i.e., approximately the same, or homoskedastic – after the rlog transformation, while they would otherwise strongly depend on the mean counts. It thus facilitates multivariate visualization and ordinations such as clustering or principal component analysis that tend to work best when the variables have similar dynamic range. Note that while the rlog transformation builds upon on our LFC shrinkage approach, it is distinct from and not part of the statistical inference

procedure for differential expression analysis described above, which employs the raw counts, not transformed data.

The rlog transformation is calculated by fitting for each gene a GLM with a baseline expression (i.e., intercept only) and, computing for each sample, shrunken LFCs with respect to the baseline, using the same empirical Bayes procedure as before (Methods). Here, however, the sample covariate information (e.g. treatment or control) is not used, so that all samples are treated equally. The rlog transformation accounts for variation in sequencing depth across samples as it represents the logarithm of q_{ij} after accounting for the size factors s_{ij}. This is in contrast to the variance-stabilizing transformation (VST) for overdispersed counts introduced in *DESeq* [4]: while the VST is also effective at stabilizing variance, it does not directly take into account differences in size factors; and in datasets with large variation in sequencing depth (dynamic range of size factors $\gtrsim$4) we observed undesirable artifacts in the performance of the VST. A disadvantage of the rlog transformation with respect to the VST is, however, that the ordering of genes within a sample will change if neighboring genes undergo shrinkage of different strength. As with the VST, the value of $\mathrm{rlog}(K_{ij})$ for large counts is approximately equal to $\log_2(K_{ij}/s_j)$. Both the rlog transformation and the VST are provided in the *DESeq2* package.

We demonstrate the use of the rlog transformation on the RNA-seq dataset of Hammer *et al.* [26], wherein RNA was sequenced from the dorsal root ganglion of rats that had undergone spinal nerve ligation and controls, at 2 weeks and at 2 months after the ligation. The count matrix for this dataset was downloaded from the ReCount online resource [27]. This dataset offers more subtle differences between conditions than the Bottomly *et al.* [16] dataset. Figure 5 provides diagnostic plots of the normalized counts under the ordinary logarithm with a pseudocount of 1 and the rlog transformation, showing that the rlog both stabilizes the variance through the range of the mean of counts and helps to find meaningful patterns in the data.

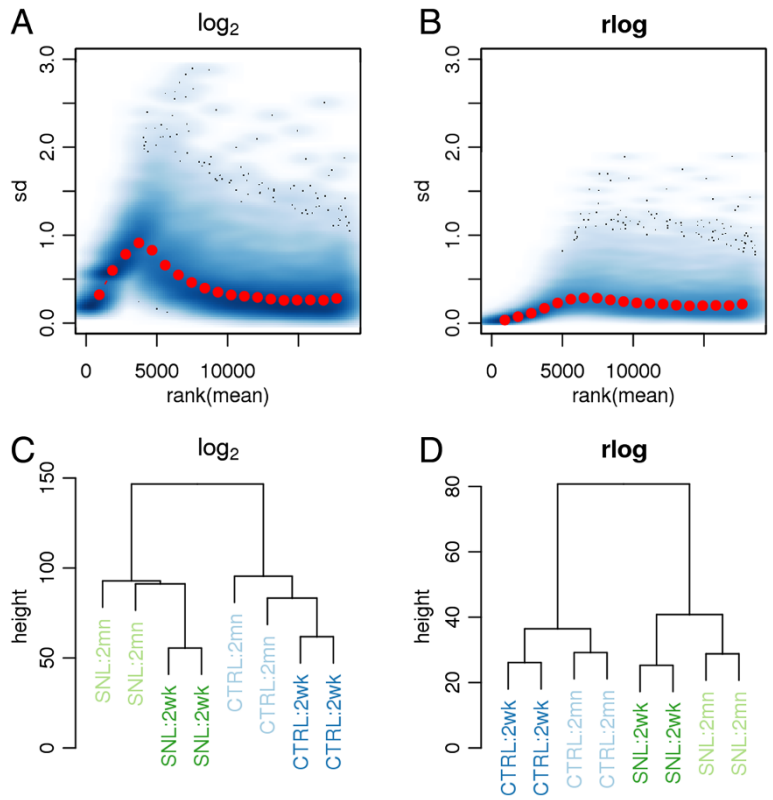


Figure 5 Variance stabilization and clustering after rlog transformation. Two transformations were applied to the counts of the Hammer *et al.* [26] dataset: the logarithm of normalized counts plus a pseudocount, i.e. $f(K_{ij}) = \log_2(K_{ij}/s_j + 1)$, and the rlog. The gene-wise standard deviation of transformed values is variable across the range of the mean of counts using the logarithm **(A)**, while relatively stable using the rlog **(B)**. A hierarchical clustering on Euclidean distances and complete linkage using the rlog **(D)** transformed data clusters the samples into the groups defined by treatment and time, while using the logarithm-transformed counts **(C)** produces a more ambiguous result. sd, standard deviation.

Gene-level analysis

We here present *DESeq2* for the analysis of per-gene counts, i.e., the total number of reads that can be uniquely assigned to a gene. In contrast, several algorithms [28,29] work with probabilistic assignments of reads to transcripts, where multiple, overlapping transcripts can originate from each gene. It has been noted that the total read count approach can result in false detection of differential expression when in fact only transcript isoform lengths change, and even in a wrong sign of LFCs in extreme cases [28]. However, in our benchmark, discussed in the following section, we found that LFC sign disagreements between total read count and probabilistic-assignment-based methods were rare for genes that were differentially expressed according to either method (Additional file 1: Figure S5). Furthermore, if estimates for average transcript length are available for the conditions, these can be incorporated into the *DESeq2* framework as gene- and sample-specific normalization factors. In addition, the approach used in *DESeq2* can be extended to isoform-specific analysis, either through generalized linear modeling at the exon level with a gene-specific mean as in the *DEXSeq* package [30] or through counting evidence for alternative isoforms in splice graphs [31,32]. In fact, the latest release version of *DEXSeq* now uses *DESeq2* as its inferential engine and so offers shrinkage estimation of dispersion and effect sizes for an exon-level analysis, too.

Comparative benchmarks

To assess how well *DESeq2* performs for standard analyses in comparison to other current methods, we used a combination of simulations and real data. The negative-binomial-based approaches compared were *DESeq (old)* [4], *edgeR* [33], *edgeR* with the robust option [34], *DSS* [6] and *EBSeq* [35]. Other methods compared were the *voom* normalization method followed by linear modeling using the *limma* package [36] and the *SAMseq* permutation method of the *samr* package [24]. For the benchmarks using real data, the *Cuffdiff 2* [28] method of the Cufflinks suite was included. For version numbers of the software used, see Additional file 1: Table S3. For all algorithms returning *P* values, the *P* values from genes with non-zero sum of read counts across samples were adjusted using the Benjamini–Hochberg procedure [21].

Benchmarks through simulation

Sensitivity and precision We simulated datasets of 10,000 genes with negative binomial distributed counts. To simulate data with realistic moments, the mean and dispersions were drawn from the joint distribution of means and gene-wise dispersion estimates from the Pickrell *et al.* data, fitting only an intercept term. These datasets were of varying total sample size ($m \in \{6, 8, 10, 20\}$), and the samples were split into two equal-sized groups; 80% of the simulated genes had no true differential expression, while for 20% of the genes, true fold changes of 2, 3 and 4 were used to generate counts across the two groups, with the direction of fold change chosen randomly. The simulated differentially expressed genes were chosen uniformly at random among all the genes, throughout the range of mean counts. MA-plots of the true fold changes used in the simulation and the observed fold changes induced by the simulation for one of the simulation settings are shown in Additional file 1: Figure S6.

Algorithms' performance in the simulation benchmark was assessed by their sensitivity and precision. The sensitivity was calculated as the fraction of genes with adjusted P value < 0.1 among the genes with true differences between group means. The precision was calculated as the fraction of genes with true differences between group means among those with adjusted P value < 0.1. The sensitivity is plotted over 1 − precision, or the FDR, in Figure 6. *DESeq2*, and also *edgeR*, often had the highest sensitivity of the algorithms that controlled type-I error in the sense that the actual FDR was at or below 0.1, the threshold for adjusted *P* values used for calling differentially expressed genes. *DESeq2* had higher sensitivity compared to the other algorithms, particularly for small fold change (2 or 3), as was also found in benchmarks performed by Zhou *et al.* [34]. For larger sample sizes and larger fold changes the performance of the various algorithms was more consistent.

The overly conservative calling of the old *DESeq* tool can be observed, with reduced sensitivity compared to the other algorithms and an actual FDR less than the nominal value of 0.1. We note that *EBSeq* version 1.4.0 by default removes low-count genes – whose 75% quantile of normalized counts is less than ten – before calling differential expression. The sensitivity of algorithms on the simulated data across a range of the mean of counts are more closely compared in Additional file 1: Figure S9.

Outlier sensitivity We used simulations to compare the sensitivity and specificity of *DESeq2*'s outlier handling approach to that of *edgeR*, which was recently added to the software and published while this manuscript was under review. *edgeR* now includes an optional method to handle outliers by iteratively refitting the GLM after down-weighting potential outlier counts [34]. The simulations, summarized in Additional file 1: Figure S10, indicated that both approaches to outliers nearly recover the performance on an outlier-free dataset, though *edgeR-robust* had slightly higher actual than nominal FDR, as seen in Additional file 1: Figure S11.

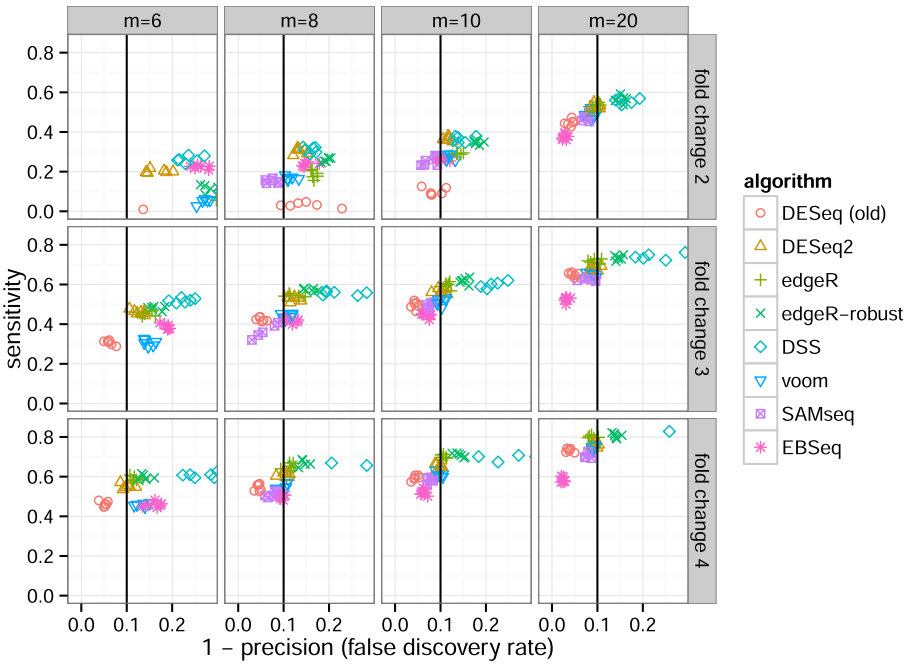


Figure 6 Sensitivity and precision of algorithms across combinations of sample size and effect size. *DESeq2* and *edgeR* often had the highest sensitivity of those algorithms that controlled the FDR, i.e., those algorithms which fall on or to the left of the vertical black line. For a plot of sensitivity against false positive rate, rather than FDR, see Additional file 1: Figure S8, and for the dependence of sensitivity on the mean of counts, see Additional file 1: Figure S9. Note that *EBSeq* filters low-count genes (see main text for details).

Precision of fold change estimates We benchmarked the *DESeq2* approach of using an empirical prior to achieve shrinkage of LFC estimates against two competing approaches: the *GFOLD* method, which can analyze experiments without replication [20] and can also handle experiments with replicates, and the *edgeR* package, which provides a pseudocount-based shrinkage termed *predictive LFCs*. Results are summarized in Additional file 1: Figures S12–S16. *DESeq2* had consistently low root-mean-square error and mean absolute error across a range of sample sizes and models for a distribution of true LFCs. *GFOLD* had similarly low error to *DESeq2* over all genes; however, when focusing on differentially expressed genes, it performed worse for larger sample sizes. *edgeR* with default settings had similarly low error to *DESeq2* when focusing only on the differentially expressed genes, but had higher error over all genes.

Clustering We compared the performance of the rlog transformation against other methods of transformation or distance calculation in the recovery of simulated clusters. The adjusted Rand index [37] was used to compare a hierarchical clustering based on various distances with the true cluster membership. We tested the Euclidean distance for normalized counts, logarithm of normalized counts plus a pseudocount of 1, rlog-transformed counts and VST counts. In addition we compared these Euclidean distances with the Poisson distance implemented in the *PoiClaClu* package [38], and a distance implemented internally in the *plotMDS* function of *edgeR* (though not the default distance, which is similar to the logarithm of normalized counts). The results, shown in Additional file 1: Figure S17, revealed that when the size factors were equal for all samples, the Poisson distance and the Euclidean distance of rlog-transformed or VST counts outperformed other methods. However, when the size factors were not equal across samples, the rlog approach generally outperformed the other methods. Finally, we note that the rlog transformation provides normalized data, which can be used for a variety of applications, of which distance calculation is one.

Benchmark for RNA sequencing data

While simulation is useful to verify how well an algorithm behaves with idealized theoretical data, and hence can verify that the algorithm performs as expected under its own assumptions, simulations cannot inform us how well the theory fits reality. With RNA-seq data, there is the complication of not knowing fully or directly the underlying truth; however, we can work around this limitation by using more indirect inference, explained below.

In the following benchmarks, we considered three performance metrics for differential expression calling: the false positive rate (or 1 minus the specificity), sensitivity

and precision. We can obtain meaningful estimates of specificity from looking at datasets where we believe all genes fall under the null hypothesis of no differential expression [39]. Sensitivity and precision are more difficult to estimate, as they require independent knowledge of those genes that are differentially expressed. To circumvent this problem, we used experimental reproducibility on independent samples (though from the same dataset) as a proxy. We used a dataset with large numbers of replicates in both of two groups, where we expect that truly differentially expressed genes exist. We repeatedly split this dataset into an evaluation set and a larger verification set, and compared the calls from the evaluation set with the calls from the verification set, which were taken as truth. It is important to keep in mind that the calls from the verification set are only an approximation of the true differential state, and the approximation error has a systematic and a stochastic component. The stochastic error becomes small once the sample size of the verification set is large enough. For the systematic errors, our benchmark assumes that these affect all algorithms more or less equally and do not markedly change the ranking of the algorithms.

False positive rate To evaluate the false positive rate of the algorithms, we considered mock comparisons from a dataset with many samples and no known condition dividing the samples into distinct groups. We used the RNA-seq data of Pickrell *et al.* [17] for lymphoblastoid cell lines derived from unrelated Nigerian individuals. We chose a set of 26 RNA-seq samples of the same read length (46 base pairs) from male individuals. We randomly drew without replacement ten samples from the set to compare five against five, and this process was repeated 30 times. We estimated the false positive rate associated with a critical value of 0.01 by dividing the number of P values less than 0.01 by the total number of tests; genes with zero sum of read counts across samples were excluded. The results over the 30 replications, summarized in Figure 7, indicated that all algorithms generally controlled the number of false positives. *DESeq (old)* and *Cuffdiff 2* appeared overly conservative in this analysis, not using up their type-I error budget.

Sensitivity To obtain an impression of the sensitivity of the algorithms, we considered the Bottomly *et al.* [16] dataset, which contains ten and eleven replicates of two different, genetically homogeneous mice strains. This allowed for a split of three vs three for the evaluation set and seven vs eight for the verification set, which were balanced across the three experimental batches. Random splits were replicated 30 times. Batch information was not provided to the *DESeq (old)*, *DESeq2*, *DSS*, *edgeR* or *voom* algorithms, which can accommodate complex

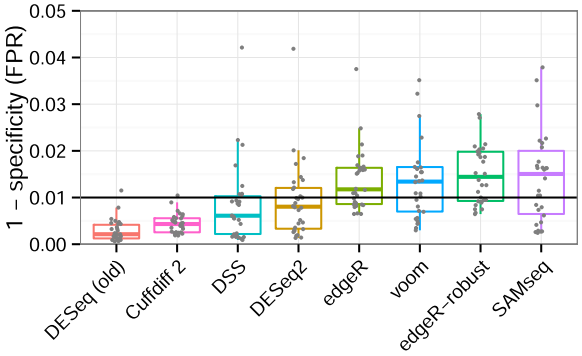


Figure 7 Benchmark of false positive calling. Shown are estimates of $P(P \text{ value} < 0.01)$ under the null hypothesis. The FPR is the number of P values less than 0.01 divided by the total number of tests, from randomly selected comparisons of five vs five samples from the Pickrell *et al.* [17] dataset, with no known condition dividing the samples. Type-I error control requires that the tool does not substantially exceed the nominal value of 0.01 (black line). *EBSeq* results were not included in this plot as it returns posterior probabilities, which unlike P values are not expected to be uniformly distributed under the null hypothesis. FPR, false positive rate.

experimental designs, to have comparable calls across all algorithms.

We rotated though each algorithm to determine the calls of the verification set. For a given algorithm's verification set calls, we tested the evaluation set calls of every algorithm. We used this approach rather than a consensus-based method, as we did not want to favor or disfavor any particular algorithm or group of algorithms. Sensitivity was calculated as in the simulation benchmark, now with true differential expression defined by an adjusted P value < 0.1 in the larger verification set, as diagrammed in Additional file 1: Figure S18. Figure 8 displays the estimates of sensitivity for each algorithm pair.

The ranking of algorithms was generally consistent regardless of which algorithm was chosen to determine calls in the verification set. *DESeq2* had comparable sensitivity to *edgeR* and *voom* though less than *DSS*. The median sensitivity estimates were typically between 0.2 and 0.4 for all algorithms. That all algorithms had relatively low median sensitivity can be explained by the small sample size of the evaluation set and the fact that increasing the sample size in the verification set increases power. It was expected that the permutation-based *SAMseq* method would rarely produce adjusted P value < 0.1 in the evaluation set, because the three vs three comparison does not enable enough permutations.

Precision Another important consideration from the perspective of an investigator is the precision, or fraction of true positives in the set of genes which pass the adjusted P value threshold. This can also be reported as

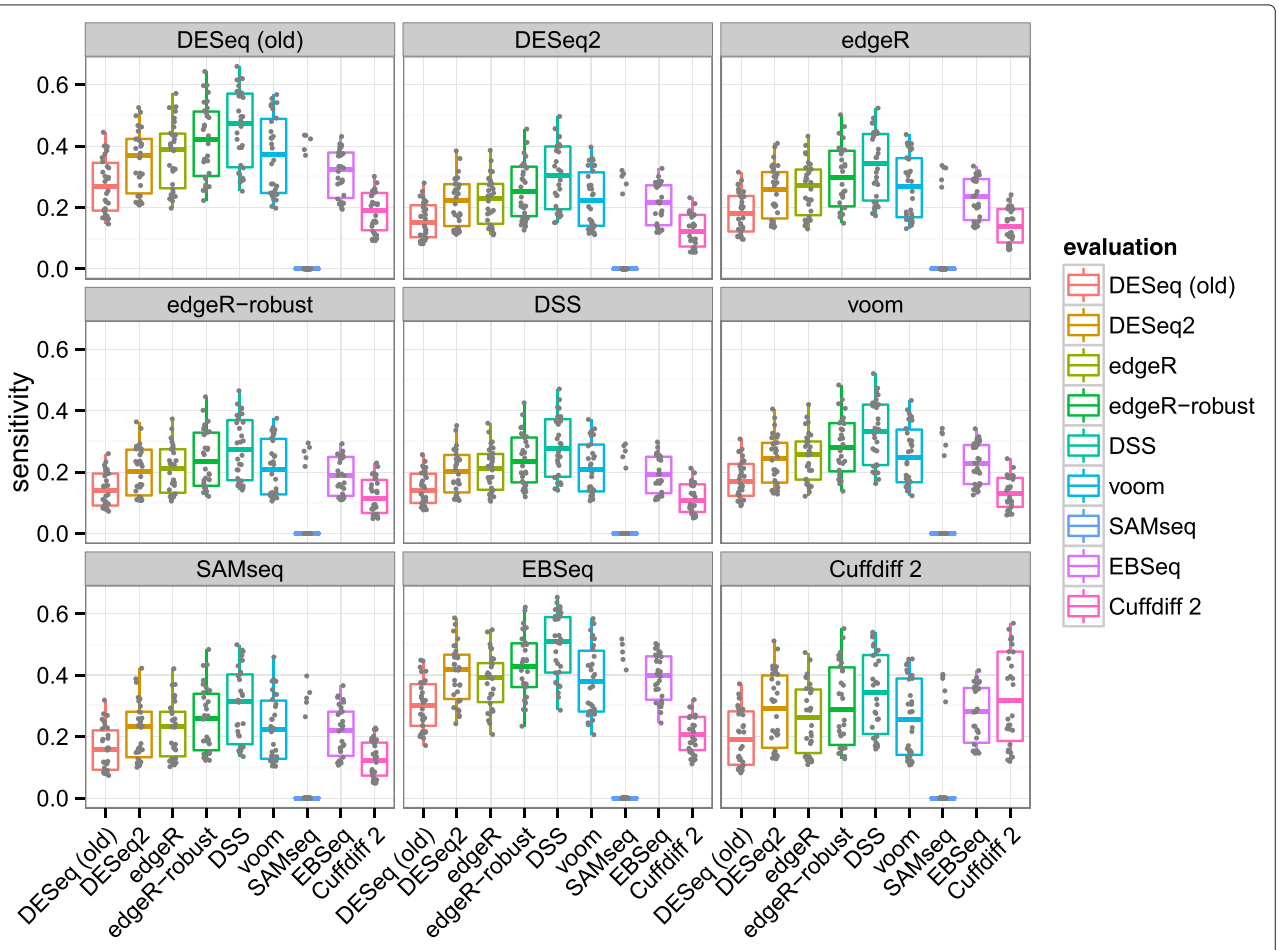


Figure 8 Sensitivity estimated from experimental reproducibility. Each algorithm's sensitivity in the evaluation set (box plots) is evaluated using the calls of each other algorithm in the verification set (panels with grey label).

1 − FDR. Again, 'true' differential expression was defined by an adjusted P value < 0.1 in the larger verification set. The estimates of precision are displayed in Figure 9, where we can see that *DESeq2* often had the second highest median precision, behind *DESeq (old)*. We can also see that algorithms with higher median sensitivity, e.g., *DSS*, were generally associated here with lower median precision. The rankings differed significantly when *Cuffdiff 2* was used to determine the verification set calls. This is likely due to the additional steps *Cuffdiff 2* performed to deconvolve changes in isoform-level abundance from gene-level abundance, which apparently came at the cost of lower precision when compared against its own verification set calls.

To compare the sensitivity and precision results further, we calculated the precision of algorithms along a grid of nominal adjusted P values (Additional file 1: Figure S19). We then found the nominal adjusted P value for each algorithm, which resulted in a median actual precision of 0.9 (FDR = 0.1). Having thus calibrated each algorithm to a target FDR, we evaluated the sensitivity of calling, as shown in Additional file 1: Figure S20. As expected, here the algorithms performed more similarly to each other. This analysis revealed that, for a given target precision, *DESeq2* often was among the top algorithms by median sensitivity, though the variability across random replicates was larger than the differences between algorithms.

The absolute number of calls for the evaluation and verification sets can be seen in Additional file 1: Figures S21 and S22, which mostly matched the order seen in the sensitivity plot of Figure 8. Additional file 1: Figures S23 and S24 provide heat maps and clustering based on the Jaccard index of calls for one replicate of the evaluation and verification sets, indicating a large overlap of calls across the different algorithms.

In summary, the benchmarking tests showed that *DESeq2* effectively controlled type-I errors, maintaining a median false positive rate just below the chosen critical value in a mock comparison of groups of samples randomly chosen from a larger pool. For both simulation and

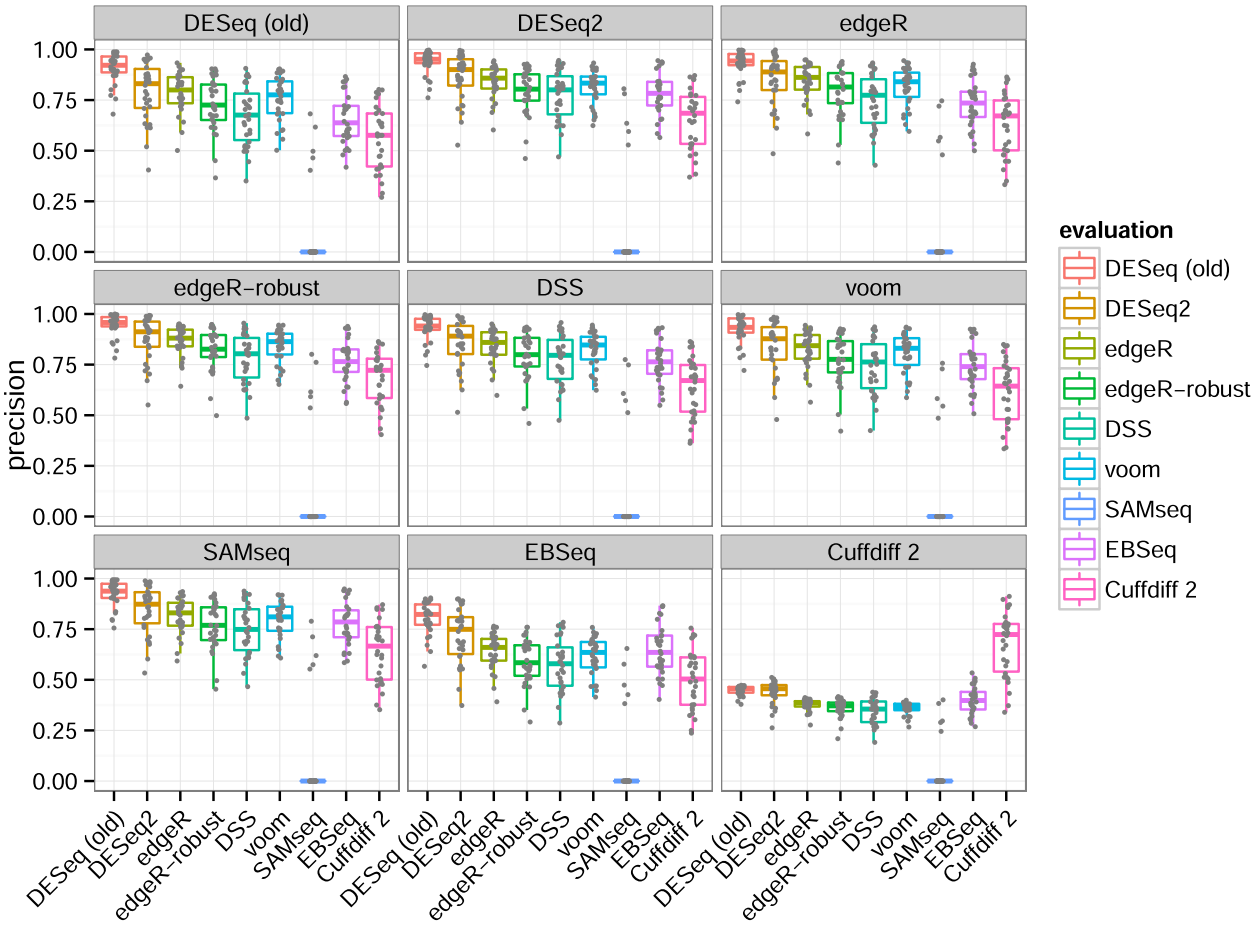


Figure 9 Precision estimated from experimental reproducibility. Each algorithm's precision in the evaluation set (box plots) is evaluated using the calls of each other algorithm in the verification set (panels with grey label).

analysis of real data, *DESeq2* often achieved the highest sensitivity of those algorithms that controlled the FDR.

Conclusions

DESeq2 offers a comprehensive and general solution for gene-level analysis of RNA-seq data. Shrinkage estimators substantially improve the stability and reproducibility of analysis results compared to maximum-likelihood-based solutions. Empirical Bayes priors provide automatic control of the amount of shrinkage based on the amount of information for the estimated quantity available in the data. This allows *DESeq2* to offer consistent performance over a large range of data types and makes it applicable for small studies with few replicates as well as for large observational studies. *DESeq2*'s heuristics for outlier detection help to recognize genes for which the modeling assumptions are unsuitable and so avoids type-I errors caused by these. The embedding of these strategies in the framework of GLMs enables the treatment of both simple and complex designs.

A critical advance is the shrinkage estimator for fold changes for differential expression analysis, which offers a sound and statistically well-founded solution to the practically relevant problem of comparing fold change across the wide dynamic range of RNA-seq experiments. This is of value for many downstream analysis tasks, including the ranking of genes for follow-up studies and association of fold changes with other variables of interest. In addition, the rlog transformation, which implements shrinkage of fold changes on a per-sample basis, facilitates visualization of differences, for example in heat maps, and enables the application of a wide range of techniques that require homoskedastic input data, including machine-learning or ordination techniques such as principal component analysis and clustering.

DESeq2 hence offers to practitioners a wide set of features with state-of-the-art inferential power. Its use cases are not limited to RNA-seq data or other transcriptomics assays; rather, many kinds of high-throughput count data can be used. Other areas for which *DESeq* or *DESeq2*

have been used include chromatin immunoprecipitation sequencing assays (e.g., [40]; see also the *DiffBind* package [41,42]), barcode-based assays (e.g., [43]), metagenomics data (e.g., [44]), ribosome profiling [45] and CRISPR/Cas-library assays [46]. Finally, the *DESeq2* package is integrated well in the Bioconductor infrastructure [11] and comes with extensive documentation, including a vignette that demonstrates a complete analysis step by step and discusses advanced use cases.

Materials and methods

A summary of the notation used in the following section is provided in Additional file 1: Table S1.

Model and normalization

The read count K_{ij} for gene i in sample j is described with a GLM of the negative binomial family with a logarithmic link:

$$\begin{aligned} K_{ij} &\sim \mathrm{NB}(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i) \\ \mu_{ij} &= s_{ij} q_{ij} \end{aligned} \tag{1}$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}. \tag{2}$$

For notational simplicity, the equations here use the natural logarithm as the link function, though the *DESeq2* software reports estimated model coefficients and their estimated standard errors on the $\log_2$ scale.

By default, the normalization constants s_{ij} are considered constant within a sample, $s_{ij} = s_j$, and are estimated with the median-of-ratios method previously described and used in *DESeq* [4] and *DEXSeq* [30]:

$$s_j = \underset{i:\, K_i^{\mathrm{R}} \neq 0}{\mathrm{median}} \frac{K_{ij}}{K_i^{\mathrm{R}}} \quad \text{with} \quad K_i^{\mathrm{R}} = \left(\prod_{j=1}^{m} K_{ij} \right)^{1/m}.$$

Alternatively, the user can supply normalization constants s_{ij} calculated using other methods (e.g., using *cqn* [13] or *EDASeq* [14]), which may differ from gene to gene.

Expanded design matrices

For consistency with our software's documentation, in the following text we will use the terminology of the *R* statistical language. In linear modeling, a categorical variable or *factor* can take on two or more values or *levels*. In standard design matrices, one of the values is chosen as a reference value or *base level* and absorbed into the intercept. In standard GLMs, the choice of base level does not influence the values of contrasts (LFCs). This, however, is no longer the case in our approach using ridge-regression-like shrinkage on the coefficients (described below), when factors with more than two levels are present in the design matrix, because the base level will not undergo shrinkage while the other levels do.

To recover the desirable symmetry between all levels, *DESeq2* uses *expanded design matrices*, which include an indicator variable for *each* level of each factor, in addition to an intercept column (i.e., none of the levels is absorbed into the intercept). While such a design matrix no longer has full rank, a unique solution exists because the zero-centered prior distribution (see below) provides regularization. For dispersion estimation and for estimating the width of the LFC prior, standard design matrices are used.

Contrasts

Contrasts between levels and standard errors of such contrasts can be calculated as they would in the standard design matrix case, i.e., using:

$$\beta_i^c = \vec{c}^{\,t} \vec{\beta}_i \tag{3}$$

$$\mathrm{SE}\left(\beta_i^c\right) = \sqrt{\vec{c}^{\,t} \Sigma_i \vec{c}}, \tag{4}$$

where $\vec{c}$ represents a numeric contrast, e.g., 1 and -1 specifying the numerator and denominator of a simple two-level contrast, and $\Sigma_i = \mathrm{Cov}(\vec{\beta}_i)$, defined below.

Estimation of dispersions

We assume the dispersion parameter α_i follows a log-normal prior distribution that is centered around a trend that depends on the gene's mean normalized read count:

$$\log \alpha_i \sim N\left(\log \alpha_{\mathrm{tr}}(\bar{\mu}_i),\, \sigma_{\mathrm{d}}^2\right). \tag{5}$$

Here, α_{tr} is a function of the gene's mean normalized count,

$$\bar{\mu}_i = \frac{1}{m} \sum_j \frac{K_{ij}}{s_{ij}}.$$

It describes the mean-dependent expectation of the prior. σ_{d} is the width of the prior, a hyperparameter describing how much the individual genes' true dispersions scatter around the trend. For the trend function, we use the same parametrization as we used for *DEXSeq* [30], namely,

$$\alpha_{\mathrm{tr}}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + \alpha_0. \tag{6}$$

We get final dispersion estimates from this model in three steps, which implement a computationally fast approximation to a full empirical Bayes treatment. We first use the count data for each gene separately to get preliminary gene-wise dispersion estimates α_i^{gw} by maximum-likelihood estimation. Then, we fit the dispersion trend α_{tr}. Finally, we combine the likelihood with the trended prior to get maximum *a posteriori* (MAP) values as final dispersion estimates. Details for the three steps follow.

Gene-wise dispersion estimates To get a gene-wise dispersion estimate for a gene i, we start by fitting a negative binomial GLM without an LFC prior for the design matrix X to the gene's count data. This GLM uses a rough method-of-moments estimate of dispersion, based on the within-group variances and means. The initial GLM is necessary to obtain an initial set of fitted values, $\hat{\mu}_{ij}^0$. We then maximize the Cox–Reid adjusted likelihood of the dispersion, conditioned on the fitted values $\hat{\mu}_{ij}^0$ from the initial fit, to obtain the gene-wise estimate α_i^{gw}, i.e.,

$$\alpha_i^{\mathrm{gw}} = \arg\max_{\alpha} \ell_{\mathrm{CR}}\left(\alpha; \vec{\mu}_{i\cdot}^0, \vec{K}_{i\cdot}\right)$$

with

$$\begin{aligned} \ell_{\mathrm{CR}}(\alpha; \vec{\mu}, \vec{K}) &= \ell(\alpha) - \frac{1}{2}\log\left(\det\left(X^t W X\right)\right) \\ \ell(\alpha) &= \sum_j \log f_{\mathrm{NB}}(K_j; \mu_j, \alpha), \end{aligned} \tag{7}$$

where $f_{\mathrm{NB}}(k; \mu, \alpha)$ is the probability mass function of the negative binomial distribution with mean μ and dispersion α, and the second term provides the Cox–Reid bias adjustment [47]. This adjustment, first used in the context of dispersion estimation for SAGE data [48] and then for HTS data [3] in *edgeR*, corrects for the negative bias of dispersion estimates from using the MLEs for the fitted values $\hat{\mu}_{ij}^0$ (analogous to Bessel's correction in the usual sample variance formula; for details, see [49], Section 10.6). It is formed from the Fisher information for the fitted values, which is here calculated as $\det(X^t W X)$, where W is the diagonal weight matrix from the standard iteratively reweighted least-squares algorithm. As the GLM's link function is $g(\mu) = \log(\mu)$ and its variance function is $V(\mu; \alpha) = \mu + \alpha\mu^2$, the elements of the diagonal matrix W_i are given by:

$$w_{jj} = \frac{1}{g'(\mu_j)^2 V(\mu_j)} = \frac{1}{1/\mu_j + \alpha}.$$

The optimization in Equation (7) is performed on the scale of $\log\alpha$ using a backtracking line search with accepted proposals that satisfy Armijo conditions [50].

Dispersion trend A parametric curve of the form (6) is fit by regressing the gene-wise dispersion estimates α_i^{gw} onto the means of the normalized counts, $\bar{\mu}_i$. The sampling distribution of the gene-wise dispersion estimate around the true value α_i can be highly skewed, and therefore we do not use ordinary least-squares regression but rather gamma-family GLM regression. Furthermore, dispersion outliers could skew the fit and hence a scheme to exclude such outliers is used.

The hyperparameters a_1 and α_0 of (6) are obtained by iteratively fitting a gamma-family GLM. At each iteration, genes with a ratio of dispersion to fitted value outside the range $[\,10^{-4}, 15]$ are left out until the sum of squared LFCs of the new coefficients over the old coefficients is less than 10^{-6} (same approach as in *DEXSeq* [30]).

The parametrization (6) is based on reports by us and others of decreasing dependence of dispersion on the mean in many datasets [3-6,51]. Some caution is warranted to disentangle true underlying dependence from effects of estimation bias that can create a perceived dependence of the dispersion on the mean. Consider a negative binomial distributed random variable with expectation μ and dispersion α. Its variance $v = \mu + \alpha\mu^2$ has two components, $v = v_{\mathrm{P}} + v_{\mathrm{D}}$, the Poisson component $v_{\mathrm{P}} = \mu$ independent of α, and the overdispersion component $v_{\mathrm{D}} = \alpha\mu^2$. When μ is small, $\mu \lesssim 1/\alpha$ (vertical lines in Additional file 1: Figure S1), the Poisson component dominates, in the sense that $v_{\mathrm{P}}/v_{\mathrm{D}} = 1/(\alpha\mu) \gtrsim 1$, and the observed data provide little information on the value of α. Therefore the sampling variance of an estimator for α will be large when $\mu \lesssim 1/\alpha$, which leads to the appearance of bias. For simplicity, we have stated the above argument without regard to the influence of the size factors, s_j, on the value of μ. This is permissible because, by construction, the geometric mean of our size factors is close to 1, and hence, the mean across samples of the unnormalized read counts, $\frac{1}{m}\sum_j K_{ij}$, and the mean of the normalized read counts, $\frac{1}{m}\sum_j K_{ij}/s_j$, will be roughly the same.

This phenomenon may give rise to an apparent dependence of α on μ. It is possible that the shape of the dispersion-mean fit for the Bottomly data (Figure 1A) can be explained in that manner: the asymptotic dispersion is $\alpha_0 \approx 0.01$, and the non-zero slope of the mean-dispersion plot is limited to the range of mean counts up to around 100, the reciprocal of α_0. However, overestimation of α in that low-count range has little effect on inference, as in that range the variance v is anyway dominated by the α-independent Poisson component v_{P}. The situation is different for the Pickrell data: here, a dependence of dispersion on mean was observed for counts clearly above the reciprocal of the asymptotic dispersion α_0 (Figure 1B), and hence is not due merely to estimation bias. Simulations (shown in Additional file 1: Figure S25) confirmed that the observed joint distribution of estimated dispersions and means is not compatible with a single, constant dispersion. Therefore, the parametrization (6) is a flexible and mildly conservative modeling choice: it is able to pick up dispersion-mean dependence if it is present, while it can lead to a minor loss of power in the low-count range due to a tendency to overestimate dispersion there.

Dispersion prior As also observed by Wu *et al.* [6], a log-normal prior fits the observed dispersion distribution for typical RNA-seq datasets. We solve the computational

difficulty of working with a non-conjugate prior using the following argument: the logarithmic residuals from the trend fit, $\log \alpha_i^{\mathrm{gw}} - \log \alpha_{\mathrm{tr}}(\bar{\mu}_i)$, arise from two contributions, namely the scatter of the true logarithmic dispersions around the trend, given by the prior with variance σ_{d}^2, and the sampling distribution of the logarithm of the dispersion estimator, with variance σ_{lde}^2. The sampling distribution of a dispersion estimator is approximately a scaled χ^2 distribution with $m - p$ degrees of freedom, with m the number of samples and p the number of coefficients. The variance of the logarithm of a χ_f^2-distributed random variable is given [52] by the trigamma function ψ_1,

$$\mathrm{Var} \log X^2 = \psi_1(f/2) \quad \text{for} \quad X^2 \sim \chi_f^2.$$

Therefore, $\sigma_{\mathrm{lde}}^2 \approx \psi_1((m-p)/2)$, i.e., the sampling variance of the logarithm of a variance or dispersion estimator is approximately constant across genes and depends only on the degrees of freedom of the model.

Additional file 1: Table S2 compares this approximation for the variance of logarithmic dispersion estimates with the variance of logarithmic Cox–Reid adjusted dispersion estimates for simulated negative binomial data, over a combination of different sample sizes, number of parameters and dispersion values used to create the simulated data. The approximation is close to the sample variance for various typical values of m, p and α.

Therefore, the prior variance σ_{d}^2 is obtained by subtracting the expected sampling variance from an estimate of the variance of the logarithmic residuals, s_{lr}^2:

$$\sigma_{\mathrm{d}}^2 = \max\left\{s_{\mathrm{lr}}^2 - \psi_1((m-p)/2),\ 0.25\right\}.$$

The prior variance σ_{d}^2 is thresholded at a minimal value of 0.25 so that the dispersion estimates are not shrunk entirely to $\alpha_{\mathrm{tr}}(\bar{\mu}_i)$ if the variance of the logarithmic residuals is less than the expected sampling variance.

To avoid inflation of σ_{d}^2 due to dispersion outliers (i.e., genes not well captured by this prior; see below), we use a robust estimator for the standard deviation s_{lr} of the logarithmic residuals,

$$s_{\mathrm{lr}} = \underset{i}{\mathrm{mad}}\left(\log \alpha_i^{\mathrm{gw}} - \log \alpha_{\mathrm{tr}}(\bar{\mu}_i)\right), \tag{8}$$

where mad stands for the median absolute deviation, divided as usual by the scaling factor $\Phi^{-1}(3/4)$.

Three or less residuals degrees of freedom When there are three or less residual degrees of freedom (number of samples minus number of parameters to estimate), the estimation of the prior variance σ_{d}^2 using the observed variance of logarithmic residuals s_{lr}^2 tends to underestimate σ_{d}^2. In this case, we instead estimate the prior variance through simulation. We match the distribution of logarithmic residuals to a density of simulated logarithmic residuals. These are the logarithm of χ_{m-p}^2-distributed random variables added to $N(0, \sigma_{\mathrm{d}}^2)$ random variables to account for the spread due to the prior. The simulated distribution is shifted by $-\log(m-p)$ to account for the scaling of the χ^2 distribution. We repeat the simulation over a grid of values for σ_{d}^2, and select the value that minimizes the Kullback–Leibler divergence from the observed density of logarithmic residuals to the simulated density.

Final dispersion estimate We form a logarithmic posterior for the dispersion from the Cox–Reid adjusted logarithmic likelihood (7) and the logarithmic prior (5) and use its maximum (i.e., the MAP value) as the final estimate of the dispersion,

$$\alpha_i^{\mathrm{MAP}} = \arg\max_{\alpha}\left(\ell_{\mathrm{CR}}\left(\alpha; \vec{\mu}_{i\cdot}^{0}, \vec{K}_{i\cdot}\right) + \Lambda_i(\alpha)\right), \tag{9}$$

where

$$\Lambda_i(\alpha) = \frac{-\left(\log\alpha - \log\alpha_{\mathrm{tr}}(\bar{\mu}_i)\right)^2}{2\sigma_{\mathrm{d}}^2},$$

is, up to an additive constant, the logarithm of the density of prior (5). Again, a backtracking line search is used to perform the optimization.

Dispersion outliers For some genes, the gene-wise estimate α_i^{gw} can be so far above the prior expectation $\alpha_{\mathrm{tr}}(\bar{\mu}_i)$ that it would be unreasonable to assume that the prior is suitable for the gene. If the dispersion estimate for such genes were down-moderated toward the fitted trend, this might lead to false positives. Therefore, we use the heuristic of considering a gene as a dispersion outlier, if the residual from the trend fit is more than two standard deviations of logarithmic residuals, s_{lr} (see Equation (8)), above the fit, i.e., if

$$\log \alpha_i^{\mathrm{gw}} > \log \alpha_{\mathrm{tr}}(\bar{\mu}_i) + 2 s_{\mathrm{lr}}.$$

For such genes, the gene-wise estimate α_i^{gw} is not shrunk toward the trended prior mean. Instead of the MAP value α_i^{MAP}, we use the gene-wise estimate α_i^{gw} as a final dispersion value in the subsequent steps. In addition, the iterative fitting procedure for the parametric dispersion trend described above avoids that such dispersion outliers influence the prior mean.

Shrinkage estimation of logarithmic fold changes

To incorporate empirical Bayes shrinkage of LFCs, we postulate a zero-centered normal prior for the coefficients β_{ir} of model (2) that represent LFCs (i.e., typically, all coefficients except for the intercept β_{i0}):

$$\beta_{ir} \sim N\left(0, \sigma_r^2\right). \tag{10}$$

As was observed with differential expression analysis using microarrays, genes with low intensity values tend to suffer from a small signal-to-noise ratio. Alternative estimators can be found that are more stable than the standard calculation of fold change as the ratio of average observed values for each condition [53-55]. *DESeq2*'s approach can be seen as an extension of these approaches for stable estimation of gene-expression fold changes to count data.

Empirical prior estimate To obtain values for the empirical prior widths σ_r for the model coefficients, we again approximate a full empirical Bayes approach, as with the estimation of dispersion prior, though here we do not subtract the expected sampling variance from the observed variance of maximum likelihood estimates. The estimate of the LFC prior width is calculated as follows. We use the standard iteratively reweighted least-squares algorithm [12] for each gene's model, Equations (1) and (2), to get MLEs for the coefficients $\beta_{ir}^{\mathrm{MLE}}$. We then fit, for each column r of the design matrix (except for the intercept), a zero-centered normal distribution to the empirical distribution of MLE fold change estimates $\vec{\beta}_r^{\mathrm{MLE}}$.

To make the fit robust against outliers with very high absolute LFC values, we use quantile matching: the width σ_r is chosen such that the $(1-p)$ empirical quantile of the absolute value of the observed LFCs, $\vec{\beta}_r^{\mathrm{MLE}}$, matches the $(1-p/2)$ theoretical quantile of the prior, $N(0,\sigma_r^2)$, where p is set by default to 0.05. If we write the theoretical upper quantile of a normal distribution as $Q_N(1-p)$ and the empirical upper quantile of the MLE LFCs as $Q_{|\beta_r|}(1-p)$, then the prior width is calculated as:

$$\sigma_r = \frac{Q_{|\beta_r|}(1-p)}{Q_N(1-p/2)}.$$

To ensure that the prior width σ_r will be independent of the choice of base level, the estimates from the quantile matching procedure are averaged for each factor over all possible contrasts of factor levels. When determining the empirical upper quantile, extreme LFC values ($|\beta_{ir}^{\mathrm{MLE}}| >$ log(2) 10, or 10 on the base 2 scale) are excluded.

Final estimate of logarithmic fold changes The logarithmic posterior for the vector, $\vec{\beta}_i$, of model coefficients β_{ir} for gene i is the sum of the logarithmic likelihood of the GLM (2) and the logarithm of the prior density (10), and its maximum yields the final MAP coefficient estimates:

$$\vec{\beta}_i = \arg\max_{\vec{\beta}} \left(\sum_j \log f_{\mathrm{NB}}\left(K_{ij}; \mu_j(\vec{\beta}), \alpha_i\right) + \Lambda(\vec{\beta}) \right),$$

where

$$\mu_j(\vec{\beta}) = s_{ij} e^{\sum_r x_{jr}\beta_r}, \quad \Lambda(\vec{\beta}) = \sum_r \frac{-\beta_r^2}{2\sigma_r^2},$$

and α_i is the final dispersion estimate for gene i, i.e., $\alpha_i = \alpha_i^{\mathrm{MAP}}$, except for dispersion outliers, where $\alpha_i = \alpha_i^{\mathrm{gw}}$.

The term $\Lambda(\beta)$, i.e., the logarithm of the density of the normal prior (up to an additive constant), can be read as a ridge penalty term, and therefore, we perform the optimization using the *iteratively reweighted ridge regression algorithm* [56], also known as *weighted updates* [57]. Specifically, the updates for a given gene are of the form

$$\vec{\beta} \leftarrow \left(X^t W X + \vec{\lambda} I\right)^{-1} X^t W \vec{z},$$

with $\lambda_r = 1/\sigma_r^2$ and

$$z_j = \log\frac{\mu_j}{s_j} + \frac{K_j - \mu_j}{\mu_j},$$

where the current fitted values $\mu_j = s_j e^{\sum_r x_{jr}\beta_r}$ are computed from the current estimates $\vec{\beta}$ in each iteration.

Fisher information. The effect of the zero-centered normal prior can be understood as shrinking the MAP LFC estimates based on the amount of information the experiment provides for this coefficient, and we briefly elaborate on this here. Specifically, for a given gene i, the shrinkage for an LFC β_{ir} depends on the *observed Fisher information*, given by

$$\mathcal{J}_m(\hat{\beta}_{ir}) = -\left[\frac{\partial^2}{\partial\beta_{ir}^2}\ell\left(\vec{\beta}_i; \vec{K}_i, \alpha_i\right)\right]_{\beta_{ir}=\hat{\beta}_{ir}},$$

where $\ell\left(\vec{\beta}_i; \vec{K}_i, \alpha_i\right)$ is the logarithm of the likelihood, and partial derivatives are taken with respect to LFC β_{ir}. For a negative binomial GLM, the observed Fisher information, or peakedness of the logarithm of the profile likelihood, is influenced by a number of factors including the degrees of freedom, the estimated mean counts μ_{ij}, and the gene's dispersion estimate α_i. The prior influences the MAP estimate when the density of the likelihood and the prior are multiplied to calculate the posterior. Genes with low estimated mean values μ_{ij} or high dispersion estimates α_i have flatter profile likelihoods, as do datasets with few residual degrees of freedom, and therefore in these cases the zero-centered prior pulls the MAP estimate from a high-uncertainty MLE closer toward zero.

Wald test

The Wald test compares the beta estimate β_{ir} divided by its estimated standard error $\mathrm{SE}(\beta_{ir})$ to a standard normal distribution. The estimated standard errors are the square root of the diagonal elements of the estimated covariance matrix, Σ_i, for the coefficients, i.e., $\mathrm{SE}(\beta_{ir}) = \sqrt{\Sigma_{i,rr}}$. Contrasts of coefficients are tested similarly by forming a Wald statistics using (3) and (4). We use the following

formula for the coefficient covariance matrix for a GLM with normal prior on coefficients [56,58]:

$$\Sigma_i = \mathrm{Cov}(\vec{\beta}_i) = (X^t W X + \vec{\lambda} I)^{-1} (X^t W X) (X^t W X + \vec{\lambda} I)^{-1}.$$

The tail integrals of the standard normal distribution are multiplied by 2 to achieve a two-tailed test. The Wald test *P* values from the subset of genes that pass the independent filtering step are adjusted for multiple testing using the procedure of Benjamini and Hochberg [21].

Independent filtering

Independent filtering does not compromise type-I error control as long as the distribution of the test statistic is marginally independent of the filter statistic *under the null hypothesis* [22], and we argue in the following that this is the case in our application. The filter statistic in *DESeq2* is the mean of normalized counts for a gene, while the test statistic is p, the *P* value from the Wald test. We first consider the case where the size factors are equal and where the gene-wise dispersion estimates are used for each gene, i.e. without dispersion shrinkage. The distribution family for the negative binomial is parameterized by $\theta = (\mu, \alpha)$. Aside from discreteness of p due to low counts, for a given μ, the distribution of p is Uniform$(0, 1)$ under the null hypothesis, so p is an ancillary statistic. The sample mean of counts for gene i, $\bar{K}_i$, is boundedly complete sufficient for μ. Then from Basu's theorem, $\bar{K}_i$ and p are independent.

While for very low counts, one can observe discreteness and non-uniformity of p under the null hypothesis, *DESeq2* does not use the distribution of p in its estimation procedure – for example, *DESeq2* does not estimate the proportion of null genes using the distribution of p – so this kind of dependence of p on μ does not lead to increased type-I error.

If the size factors are not equal across samples, but not correlated with condition, conditioning on the mean of *normalized* counts should also provide uniformly distributed p as with conditioning on the mean of counts, $\bar{K}_i$. We may consider a pathological case where the size factors are perfectly confounded with condition, in which case, even under the null hypothesis, genes with low mean count would have non-uniform distribution of p, as one condition could have positive counts and the other condition often zero counts. This could lead to non-uniformity of p under the null hypothesis; however, such a pathological case would pose problems for many statistical tests of differences in mean.

We used simulation to demonstrate that the independence of the null distribution of the test statistic from the filter statistic still holds for dispersion shrinkage. Additional file 1: Figure S26 displays marginal null distributions of p across the range of mean normalized counts. Despite spikes in the distribution for the genes with the lowest mean counts due to discreteness of the data, these densities were nearly uniform across the range of average expression strength.

Composite null hypotheses

DESeq2 offers tests for composite null hypotheses of the form $\mathcal{H}_0 : |\beta_{ir}| \leq \theta$ to find genes whose LFC significantly exceeds a threshold $\theta > 0$. The composite null hypothesis is replaced by two simple null hypotheses: $\mathcal{H}_{0a} : \beta_{ir} = \theta$ and $\mathcal{H}_{0b} : \beta_{ir} = -\theta$. Two-tailed *P* values are generated by integrating a normal distribution centered on θ with standard deviation $\mathrm{SE}(\beta_{ir})$ from $|\beta_{ir}|$ toward ∞. The value of the integral is then multiplied by 2 and thresholded at 1. This procedure controls type-I error even when $\beta_{ir} = \pm\theta$, and is equivalent to the standard *DESeq2* *P* value when $\theta = 0$.

Conversely, when searching for genes whose absolute LFC is significantly below a threshold, i.e., when testing the null hypothesis $\mathcal{H}_0 : |\beta_{ir}| \geq \theta$, the *P* value is constructed as the maximum of two one-sided tests of the simple null hypotheses: $\mathcal{H}_{0a} : \beta_{ir} = \theta$ and $\mathcal{H}_{0b} : \beta_{ir} = -\theta$. The one-sided *P* values are generated by integrating a normal distribution centered on θ with standard deviation $\mathrm{SE}(\beta_{ir})$ from β_{ir} toward $-\infty$, and integrating a normal distribution centered on $-\theta$ with standard deviation $\mathrm{SE}(\beta_{ir})$ from β_{ir} toward ∞.

Note that while a zero-centered prior on LFCs is consistent with testing the null hypothesis of small LFCs, it should not be used when testing the null hypothesis of large LFCs, because the prior would then favor the alternative hypothesis. *DESeq2* requires that no prior has been used when testing the null hypothesis of large LFCs, so that the data alone must provide evidence against the null hypothesis.

Interactions

Two exceptions to the default *DESeq2* LFC estimation steps are used for experimental designs with interaction terms. First, when any interaction terms are included in the design, the LFC prior width for main effect terms is not estimated from the data, but set to a wide value ($\sigma_r^2 = (\log(2))^2\, 1000$, or 1000 on the base 2 scale). This ensures that shrinkage of main effect terms will not result in false positive calls of significance for interactions. Second, when interaction terms are included and all factors have two levels, then standard design matrices are used rather than expanded model matrices, such that only a single term is used to test the null hypothesis that a combination of two effects is merely additive in the logarithmic scale.

Regularized logarithm

The rlog transformation is calculated as follows. The experimental design matrix X is substituted with a design

matrix with an indicator variable for every sample in addition to an intercept column. A model as described in Equations (1) and (2) is fit with a zero-centered normal prior on the non-intercept terms and using the fitted dispersion values $\alpha_{\mathrm{tr}}(\bar{\mu})$, which capture the overall variance-mean dependence of the dataset. The true experimental design matrix X is then only used in estimating the variance-mean trend over all genes. For unsupervised analyses, for instance sample quality assessment, it is desirable that the experimental design has no influence on the transformation, and hence *DESeq2* by default ignores the design matrix and re-estimates the dispersions treating all samples as replicates, i.e., it uses *blind* dispersion estimation. The rlog-transformed values are the fitted values,

$$\mathrm{rlog}\left(K_{ij}\right) \equiv \log_2 q_{ij} = \beta_{i0} + \beta_{ij},$$

where β_{ij} is the shrunken LFC on the base 2 scale for the jth sample. The variance of the prior is set using a similar approach as taken with differential expression, by matching a zero-centered normal distribution to observed LFCs. First a matrix of LFCs is calculated by taking the logarithm (base 2) of the normalized counts plus a pseudocount of $\frac{1}{2}$ for each sample divided by the mean of normalized counts plus a pseudocount of $\frac{1}{2}$. The pseudocount of $\frac{1}{2}$ allows for calculation of the logarithmic ratio for all genes, and has little effect on the estimate of the variance of the prior or the final rlog transformation. This matrix of LFCs then represents the common-scale logarithmic ratio of each sample to the fitted value using only an intercept. The prior variance is found by matching the 97.5% quantile of a zero-centered normal distribution to the 95% quantile of the absolute values in the LFC matrix.

Cook's distance for outlier detection

The MLE of $\vec{\beta}_i$ is used for calculating Cook's distance. Considering a gene i and sample j, Cook's distance for GLMs is given by [59]:

$$D_{ij} = \frac{R_{ij}^2}{\tau p} \frac{h_{jj}}{(1-h_{jj})^2},$$

where R_{ij} is the Pearson residual of sample j, τ is an overdispersion parameter (in the negative binomial GLM, τ is set to 1), p is the number of parameters including the intercept, and h_{jj} is the jth diagonal element of the hat matrix H:

$$H = W^{1/2} X (X^t W X)^{-1} X^t W^{1/2}.$$

Pearson residuals R_{ij} are calculated as

$$R_{ij} = \frac{(K_{ij} - \mu_{ij})}{\sqrt{V(\mu_{ij})}},$$

where μ_{ij} is estimated by the negative binomial GLM without the LFC prior, and using the variance function $V(\mu) = \mu + \alpha\mu^2$. A method-of-moments estimate α_i^{rob}, using a robust estimator of variance $s_{i,\mathrm{rob}}^2$ to provide robustness against outliers, is used here:

$$\alpha_i^{\mathrm{rob}} = \max\left(\frac{s_{i,\mathrm{rob}}^2 - \bar{\mu}_i}{\bar{\mu}_i^2}, 0\right).$$

R/Bioconductor package

DESeq2 is implemented as a package for the R statistical environment and is available [10] as part of the Bioconductor project [11]. The count matrix and metadata, including the gene model and sample information, are stored in an S4 class derived from the *SummarizedExperiment* class of the *GenomicRanges* package [60]. *SummarizedExperiment* objects containing count matrices can be easily generated using the *summarizeOverlaps* function of the *GenomicAlignments* package [61]. This workflow automatically stores the gene model as metadata and additionally other information such as the genome and gene annotation versions. Other methods to obtain count matrices include the *htseq-count* script [62] and the Bioconductor packages *easyRNASeq* [63] and *featureCount* [64].

The *DESeq2* package comes with a detailed vignette, which works through a number of example differential expression analyses on real datasets, and the use of the rlog transformation for quality assessment and visualization. A single function, called *DESeq*, is used to run the default analysis, while lower-level functions are also available for advanced users.

Read alignment for the Bottomly *et al.* and Pickrell *et al.* datasets

Reads were aligned using the TopHat2 aligner [65], and assigned to genes using the *summarizeOverlaps* function of the *GenomicRanges* package [60]. The sequence read archive fastq files of the Pickrell *et al.* [17] dataset (accession number [SRA:SRP001540]) were aligned to the *Homo sapiens* reference sequence GRCh37 downloaded in March 2013 from Illumina iGenomes. Reads were counted in the genes defined by the Ensembl GTF file, release 70, contained in the Illumina iGenome. The sequence read archive fastq files of the Bottomly *et al.* [16] dataset (accession number [SRA:SRP004777]) were aligned to the *Mus musculus* reference sequence NCBIM37 downloaded in March 2013 from Illumina iGenomes. Reads were counted in the genes defined by

the Ensembl GTF file, release 66, contained in the Illumina iGenome.

Reproducible code

Sweave vignettes for reproducing all figures and tables in this paper, including data objects for the experiments mentioned, and code for aligning reads and for benchmarking, can be found in a package *DESeq2paper* [66].

Additional file

Additional file 1: Supplementary methods, tables and figures.

Abbreviations

FDR: False discovery rate; GLM: Generalized linear model; HTS: High-throughput sequencing; LFC: Logarithmic fold change; MAP: Maximum *a posteriori*; MLE: Maximum-likelihood estimate; RNA-seq: RNA sequencing; VST: Variance-stabilizing transformation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors developed the method and wrote the manuscript. MIL implemented the method and performed the analyses. All authors read and approved the final manuscript.

Acknowledgements

The authors thank all users of DESeq and DESeq2 who provided valuable feedback. We thank Judith Zaugg for helpful comments on the manuscript. MIL acknowledges funding via a stipend from the International Max Planck Research School for Computational Biology and Scientific Computing and a grant from the National Institutes of Health (5T32CA009337-33). WH and SA acknowledge funding from the European Union's 7th Framework Programme (Health) via Project *Radiant*. We thank an anonymous reviewer for raising the question of estimation biases in the dispersion-mean trend fitting.

Author details

[1]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute and Department of Biostatistics, Harvard School of Public Health, 450 Brookline Avenue, Boston, MA 02215, USA. [2]Genome Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

Received: 27 May 2014 Accepted: 19 November 2014
Published online: 05 December 2014

References

1. Lönnstedt I, Speed T: **Replicated microarray data.** *Stat Sinica* 2002, **12:**31–46.
2. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23:**2881–2887.
3. McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation.** *Nucleic Acids Res* 2012, **40:**4288–4297.
4. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**106.
5. Zhou Y-H, Xia K, Wright FA: **A powerful and flexible approach to the analysis of RNA sequence count data.** *Bioinformatics* 2011, **27:**2672–2678.
6. Wu H, Wang C, Wu Z: **A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data.** *Biostatistics* 2013, **14:**232–243.
7. Hardcastle T, Kelly K: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinformatics* 2010, **11:**422.
8. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN: **Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors.** *Biostatistics* 2013, **14:**113–128.
9. Boer JM, Huber WK, Sültmann H, Wilmer F, von Heydebreck A, Haas S, Korn B, Gunawan B, Vente A, Füzesi L, Vingron M, Poustka A: **Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array.** *Genome Res* 2001, **11:**1861–1870.
10. **DESeq2.** [http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html]
11. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
12. McCullagh P, Nelder JA: **Generalized linear models.** In *Monographs on Statistics & Applied Probability*. 2nd edition. London, UK: Chapman & Hall/CRC; 1989.
13. Hansen KD, Irizarry RA, Wu Z: **Removing technical variability in RNA-seq data using conditional quantile normalization.** *Biostatistics* 2012, **13:**204–216.
14. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-seq data.** *BMC Bioinformatics* 2011, **12:**480.
15. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**1–25.
16. Bottomly D, Walter NAR, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R: **Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays.** *PLoS ONE* 2011, **6:**17820.
17. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768–772.
18. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition. New York City, USA: Springer; 2009.
19. Bi Y, Davuluri R: **NPEBseq: nonparametric empirical Bayesian-based procedure for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14:**262.
20. Feng J, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y: **GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data.** *Bioinformatics* 2012, **28:**2782–2788.
21. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B Methodol* 1995, **57:**289–300.
22. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments.** *Proc Natl Acad Sci USA* 2010, **107:**9546–9551.
23. McCarthy DJ, Smyth GK: **Testing significance relative to a fold-change threshold is a TREAT.** *Bioinformatics* 2009, **25:**765–771.
24. Li J, Tibshirani R: **Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data.** *Stat Methods Med Res* 2013, **22:**519–536.
25. Cook RD: **Detection of influential observation in linear regression.** *Technometrics* 1977, **19:**15–18.
26. Hammer P, Banck MS, Amberg R, Wang C, Petznick G, Luo S, Khrebtukova I, Schroth GP, Beyerlein P, Beutler AS: **mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain.** *Genome Res* 2010, **20:**847–860.
27. Frazee A, Langmead B, Leek J: **ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets.** *BMC Bioinformatics* 2011, **12:**449.
28. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2012, **31:**46–53.
29. Glaus P, Honkela A, Rattray M: **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics* 2012, **28:**1721–1728.

30. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res* 2012, **22:**2008–2017.
31. Sammeth M: **Complete alternative splicing events are bubbles in splicing graphs.** *J Comput Biol* 2009, **16:**1117–1140.
32. Pagès H, Bindreither D, Carlson M, Morgan M: **SplicingGraphs: create, manipulate, visualize splicing graphs, and assign RNA-seq reads to them.** 2013. Bioconductor package [http://www.bioconductor.org]
33. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2009, **26:**139–140.
34. Zhou X, Lindsay H, Robinson MD: **Robustly detecting differential expression in RNA sequencing data using observation weights.** *Nucleic Acids Res* 2014, **42:**e91.
35. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C: **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.** *Bioinformatics* 2013, **29:**1035–1043.
36. Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15:**29.
37. Hubert L, Arabie P: **Comparing partitions.** *J Classif* 1985, **2:**193–218.
38. Witten DM: **Classification and clustering of sequencing data using a Poisson model.** *Ann Appl Stat* 2011, **5:**2493–2518.
39. Irizarry RA, Wu Z, Jaffee HA: **Comparison of affymetrix GeneChip expression measures.** *Bioinformatics* 2006, **22:**789–794.
40. Asangani IA, Dommeti VL, Wang X, Malik R, Cieslik M, Yang R, Escara-Wilke J, Wilder-Romans K, Dhanireddy S, Engelke C, Iyer MK, Jing X, Wu Y-M, Cao X, Qin ZS, Wang S, Feng FY, Chinnaiyan AM: **Therapeutic targeting of BET bromodomain proteins in castration-resistant prostate cancer.** *Nature* 2014, **510:**278–282.
41. Stark R, Brown G: **DiffBind: differential binding analysis of ChIP-seq peak data.** 2013. Bioconductor package [http://www.bioconductor.org]
42. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, Ali S, Chin S-F, Palmieri C, Caldas C, Carroll JS: **Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.** *Nature* 2012, **481:**389–393.
43. Robinson DG, Chen W, Storey JD, Gresham D: **Design and analysis of bar-seq experiments.** *G3 (Bethesda)* 2013, **4:**11–18.
44. McMurdie PJ, Holmes S: **Waste not, want not: why rarefying microbiome data is inadmissible.** *PLoS Comput Biol* 2014, **10:**1003531.
45. Vasquez J, Hon C, Vanselow JT, Schlosser A, Siegel TN: **Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages.** *Nucleic Acids Res* 2014, **42:**3623–3637.
46. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W: **High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells.** *Nature* 2014, **509:**487–491.
47. Cox DR, Reid N: **Parameter orthogonality and approximate conditional inference.** *J R Stat Soc Ser B Methodol* 1987, **49:**1–39.
48. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2007, **9:**321–332.
49. Pawitan Y: *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. 1st edition. New York City, USA: Oxford University Press; 2001.
50. Armijo L: **Minimization of functions having Lipschitz continuous first partial derivatives.** *Pac J Math* 1966, **16:**1–3.
51. Di Y, Schafer DW, Cumbie JS, Chang JH: **The NBP negative binomial model for assessing differential gene expression from RNA-seq.** *Stat Appl Genet Mol Biol* 2011, **10:**1–28.
52. Abramowitz M, Stegun I: *Handbook of Mathematical Functions*. 1st edition. New York, USA: Dover Publications; 1965. [Dover Books on mathematics]
53. Newton M, Kendziorski C, Richmond C, Blattner F, Tsui K: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37–52.
54. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18:**96–104.
55. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18:**105–110.
56. Park MY: **Generalized linear models with regularization.** *PhD thesis*. Stanford University, Department of Statistics; 2006.
57. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J Stat Softw* 2010, **33:**1–22.
58. Cule E, Vineis P, De Iorio M: **Significance testing in ridge regression for genetic data.** *BMC Bioinformatics* 2011, **12:**372.
59. Cook RD, Weisberg S: *Residuals and Influence in Regression*. 1st edition. New York, USA: Chapman and Hall/CRC; 1982.
60. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol* 2013, **9:**1003118.
61. Pagès H, Obenchain V, Morgan M: **GenomicAlignments: Representation and manipulation of short genomic alignments.** 2013. Bioconductor package [http://www.bioconductor.org]
62. Anders S, Pyl PT, Huber W: **HTSeq - A Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166.
63. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM: **easyRNASeq: a Bioconductor package for processing RNA-seq data.** *Bioinformatics* 2012, **28:**2532–2533.
64. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923–930.
65. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**36.
66. **DESeq2paper.** [http://www-huber.embl.de/DESeq2paper]

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Review

OPEN ACCESS

molecular systems biology

Current best practices in single-cell RNA-seq analysis: a tutorial

Malte D Luecken[1] & Fabian J Theis[1,2,*]

Abstract

Single-cell RNA-seq has enabled gene expression to be studied at an unprecedented resolution. The promise of this technology is attracting a growing user base for single-cell analysis methods. As more analysis tools are becoming available, it is becoming increasingly difficult to navigate this landscape and produce an up-to-date workflow to analyse one's data. Here, we detail the steps of a typical single-cell RNA-seq analysis, including pre-processing (quality control, normalization, data correction, feature selection, and dimensionality reduction) and cell- and gene-level downstream analysis. We formulate current best-practice recommendations for these steps based on independent comparison studies. We have integrated these best-practice recommendations into a workflow, which we apply to a public dataset to further illustrate how these steps work in practice. Our documented case study can be found at https://www.github.com/theislab/single-cell-tutorial. This review will serve as a workflow tutorial for new entrants into the field, and help established users update their analysis pipelines.

Keywords analysis pipeline development; computational biology; data analysis tutorial; single-cell RNA-seq

DOI 10.15252/msb.20188746 | Received 16 November 2018 | Revised 15 March 2019 | Accepted 3 April 2019

Mol Syst Biol. (2019) 15: e8746

Introduction

In recent years, single-cell RNA sequencing (scRNA-seq) has significantly advanced our knowledge of biological systems. We have been able to both study the cellular heterogeneity of zebrafish, frogs and planaria (Briggs *et al*, 2018; Plass *et al*, 2018; Wagner *et al*, 2018) and discover previously obscured cellular populations (Montoro *et al*, 2018; Plasschaert *et al*, 2018). The great potential of this technology has motivated computational biologists to develop a range of analysis tools (Rostom *et al*, 2017). Despite considerable effort being undertaken by the field to ensure the usability of individual tools, a barrier of entry for novices in single-cell data analysis is the lack of standardization due to the relative immaturity of the field. In this paper, we present a tutorial for scRNA-seq analysis and outline current best practices to lay a foundation for future analysis standardization.

The challenges to standardization include the growing number of analysis methods (385 tools as of 7 March 2019) and exploding dataset sizes (Angerer *et al*, 2017; Zappia *et al*, 2018). We are continuously finding new ways to use the data at our disposal. For example, it has recently become possible to predict cell fates in differentiation (La Manno *et al*, 2018). While the continuous improvement of analysis tools is beneficial for generating new scientific insight, it complicates standardization.

Further challenges for standardization lie in technical aspects. Analysis tools for scRNA-seq data are written in a variety of programming languages—most prominently R and Python (Zappia *et al*, 2018). Although cross-environment support is growing (preprint: Scholz *et al*, 2018), the choice of programming language is often also a choice between analysis tools. Popular platforms such as Seurat (Butler *et al*, 2018), Scater (McCarthy *et al*, 2017), or Scanpy (Wolf *et al*, 2018) provide integrated environments to develop pipelines and contain large analysis toolboxes. However, out of necessity these platforms limit themselves to tools developed in their respective programming languages. By extension, language restrictions also hold true for currently available scRNA-seq analysis tutorials, many of which revolve around the above platforms (R and bioconductor tools: https://github.com/drisso/bioc2016singlecell and https://hemberg-lab.github.io/scRNA.seq.course/; Lun *et al*, 2016b; Seurat: https://satijalab.org/seurat/get_started.html; Scanpy: https://scanpy.readthedocs.io/en/stable/tutorials.html).

Considering the above-mentioned challenges, instead of targeting a standardized analysis pipeline, we outline current best practices and common tools independent of programming language. We guide the reader through the various steps of a scRNA-seq analysis pipeline (Fig 1), present current best practices, and discuss analysis pitfalls and open questions. Where best practices cannot be determined due to novelty of the tools and lack of comparisons, we list popular available tools. The outlined steps start from read or count matrices and lead to potential analysis endpoints. Earlier pre-processing steps are covered in Lun et al (2016b). A detailed case study that integrates the established current best practices is available on our github at: https://github.com/theislab/single-cell-tutorial/. Here, we have applied the current best practices in a practical example workflow to analyse a public dataset. The analysis workflow

1 Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
2 Department of Mathematics, Technische Universität München, Garching bei München, Germany
*Corresponding author. Tel: +498931872211; E-mail: fabian.theis@helmholtz-muenchen.de

© 2019 The Authors. Published under the terms of the CC BY 4.0 license

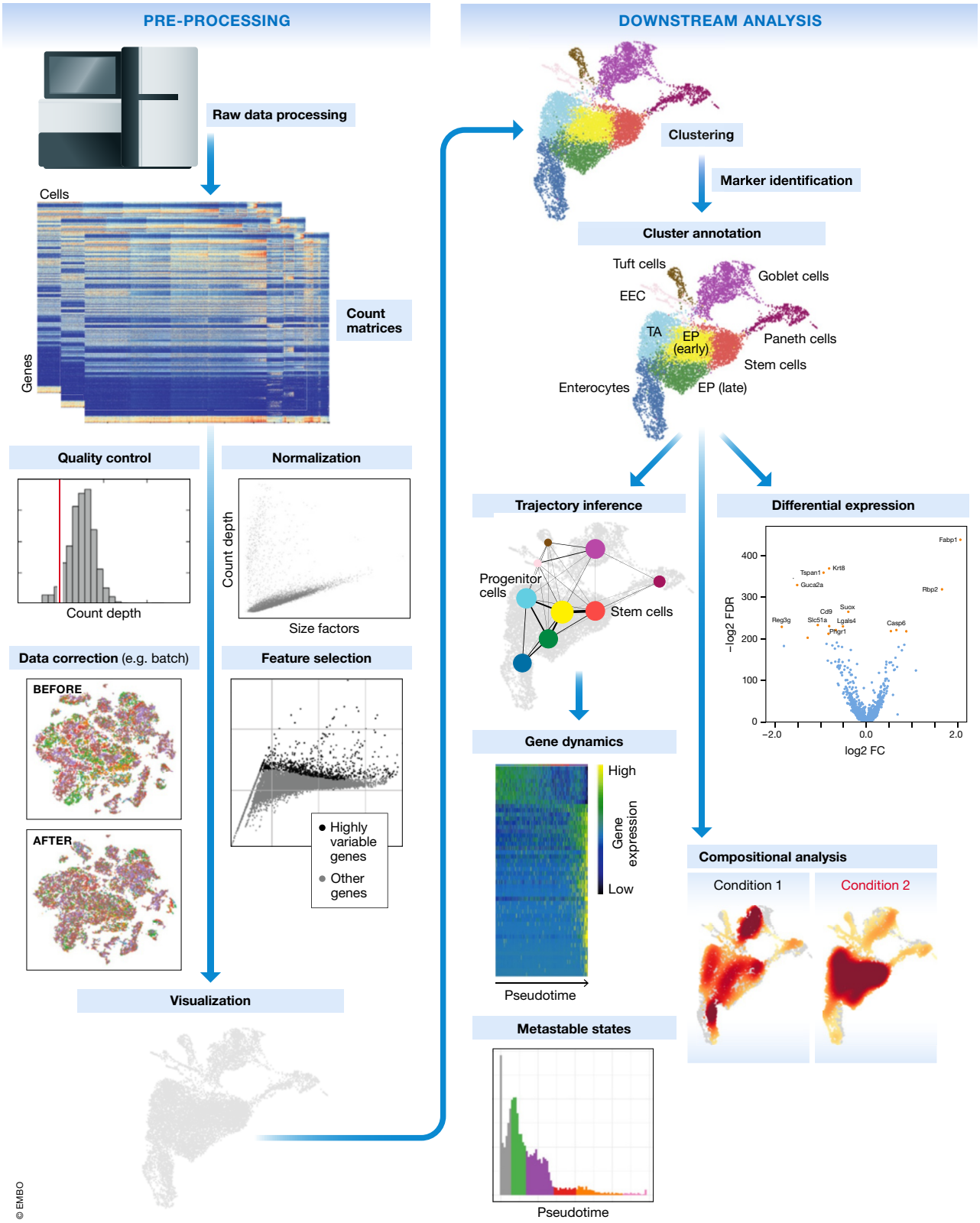


Figure 1. Schematic of a typical single-cell RNA-seq analysis workflow.
Raw sequencing data are processed and aligned to give count matrices, which represent the start of the workflow. The count data undergo pre-processing and downstream analysis. Subplots are generated using the best-practices workflow on intestinal epithelium data from Haber *et al* (2017).

 © 2019 The Authors

Box 1: Key elements of an experimental scRNA-seq workflow

Generating single-cell data from a biological sample requires multiple steps. Typical workflows incorporate single-cell dissociation, single-cell isolation, library construction, and sequencing. We give a brief overview of these stages here. A more detailed explanation and comparison of different protocols can be found in Ziegenhain *et al* (2017); Macosko *et al* (2015); Svensson *et al* (2017).

Input material for a single-cell experiment is typically obtained in the form of biological tissue samples. As a first step, a single-cell suspension is generated in a process called *single-cell dissociation* in which the tissue is digested.

To profile the mRNA in each cell separately, cells must be isolated. *Single-cell isolation* is performed differently depending on the experimental protocol. While plate-based techniques isolate cells into wells on a plate, droplet-based methods rely on capturing each cell in its own microfluidic droplet. In both cases, errors can occur that lead to multiple cells being captured together (*doublets* or *multiplets*), non-viable cells being captured, or no cell being captured at all (*empty droplets/wells*). Empty droplets are especially common as droplet-based methods rely on a low concentration flow of input cells to control doublet rates.

Each well or droplet contains the necessary chemicals to break down the cell membranes and perform *library construction*. Library construction is the process in which the intracellular mRNA is captured, reverse-transcribed to cDNA molecules and amplified. As cells undergo this process in isolation, the mRNA from each cell can be labelled with a well- or droplet-specific cellular barcode. Furthermore, many experimental protocols also label captured molecules with a *unique molecular identifier (UMI)*. Cellular cDNA is amplified before sequencing to increase its probability of being measured. UMIs allow us to distinguish between amplified copies of the same mRNA molecule and reads from separate mRNA molecules transcribed from the same gene.

After library construction, cellular cDNA libraries are labelled with cellular barcodes and, depending on the protocol, UMIs. These libraries are pooled together (*multiplexed*) for *sequencing*. Sequencing produces read data, which undergo quality control, grouping based on their assigned barcodes (*demultiplexing*) and alignment in read processing pipelines. For UMI-based protocols, read data can be further demultiplexed to produce counts of captured mRNA molecules (*count data*).

integrates R and Python tools in a Jupyter–Ipython notebook with rpy2. With the available documentation, it is readily adaptable as a workflow template.

Pre-processing and visualization

Raw data generated by sequencing machines are processed to obtain matrices of molecular counts (count matrices) or, alternatively, read counts (read matrices), depending on whether unique molecular identifiers (UMIs) were incorporated in the single-cell library construction protocol (see Box 1 for an overview of the experimental steps that precede the analysis). Raw data processing pipelines such as Cell Ranger (Zheng *et al*, 2017), indrops (Klein *et al*, 2015), SEQC (Azizi *et al*, 2018), or zUMIs (Parekh *et al*, 2018) take care of read quality control (QC), assigning reads to their cellular barcodes and mRNA molecules of origin (also called "demultiplexing"), genome alignment, and quantification. The resulting read or count matrices have the dimension number of barcodes x number of transcripts. The term "barcode" is used here instead of "cell" as all reads assigned to the same barcode may not correspond to reads from the same cell. A barcode may mistakenly tag multiple cells (doublet) or may not tag any cells (empty droplet/well).

While read and count data differ in the level of measurement noise, the processing steps in a typical analysis pipeline are the same. For simplicity, we will refer to the data as count matrices in this tutorial. Where results for read and count matrices differ, read matrices are specifically mentioned.

Quality control

Before analysing the single-cell gene expression data, we must ensure that all cellular barcode data correspond to viable cells. Cell QC is commonly performed based on three QC covariates: the number of counts per barcode (count depth), the number of genes per barcode, and the fraction of counts from mitochondrial genes per barcode (Ilicic *et al*, 2016; Griffiths *et al*, 2018). The distributions of these QC covariates are examined for outlier peaks that are filtered out by thresholding (Fig 2). These outlier barcodes can correspond to dying cells, cells whose membranes are broken, or doublets. For example, barcodes with a low count depth, few detected genes, and a high fraction of mitochondrial counts are indicative of cells whose cytoplasmic mRNA has leaked out through a broken membrane, and thus, only mRNA located in the mitochondria is still conserved (Fig 2). In contrast, cells with unexpectedly high counts and a large number of detected genes may represent doublets. Thus, high-count depth thresholds are commonly used to filter out potential doublets. Three recent doublet detection tools offer more elegant and potentially better solutions (DoubletDecon: preprint: DePasquale *et al*, 2018; Scrublet: Wolock *et al*, 2019; Doublet Finder: McGinnis *et al*, 2018).

Considering any of these three QC covariates in isolation can lead to misinterpretation of cellular signals. For example, cells with a comparatively high fraction of mitochondrial counts may be involved in respiratory processes. Likewise, other QC covariates also have biological interpretations. Cells with low counts and/or genes may correspond to quiescent cell populations, and cells with high counts may be larger in size. Indeed, molecular counts can differ strongly between cells (see case study on project github). Thus, QC covariates should be considered jointly when univariate thresholding decisions are made (Fig 2D), and these thresholds should be set as permissive as possible to avoid filtering out viable cell populations unintentionally. In future, filtering models that account for multivariate QC dependencies may provide more sensitive QC options.

Datasets that contain heterogeneous mixtures of cell types may exhibit multiple QC covariate peaks. For example, Fig 2D shows two populations of cells with different QC distributions. If no previous filtering step was performed (note that Cell Ranger also performs cell QC), then only the lowest count depth and gene per barcode peak should be considered as non-viable cells. A further thresholding guideline is the proportion of cells that are filtered out with the chosen threshold. For high-count filtering, this proportion should not exceed the expected doublet rate.

In addition to checking the integrity of cells, QC steps must also be performed at the level of transcripts. Raw count matrices often include over 20,000 genes. This number can be drastically reduced by filtering out genes that are not expressed in more than a few cells and are thus not informative of the cellular heterogeneity. A guideline to setting this threshold is to use the minimum cell cluster size that is

© 2019 The Authors

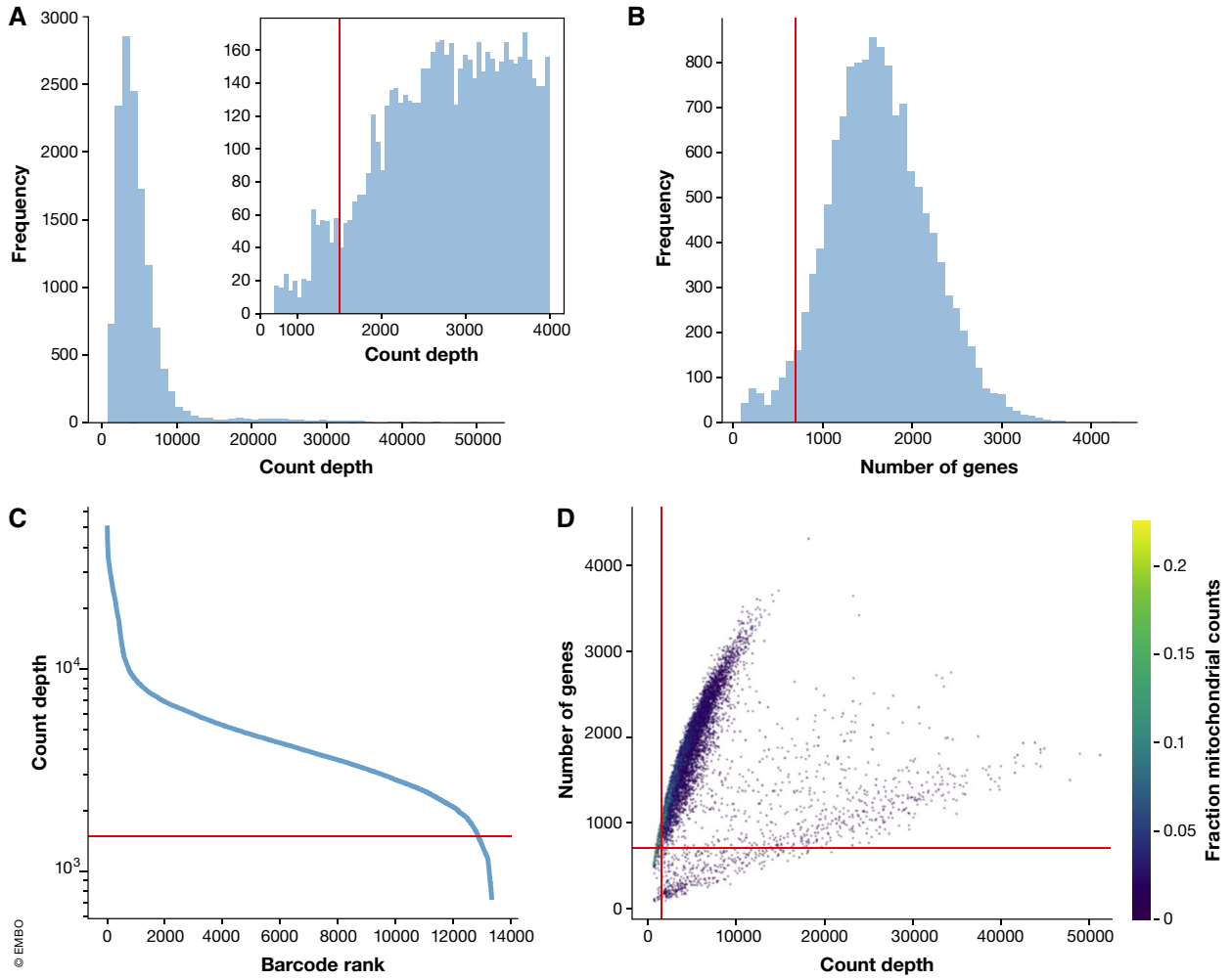


Figure 2. Plots of quality control metrics with filtering decisions for a mouse intestinal epithelium dataset from Haber *et al* (2017).

(A) Histograms of count depth per cell. The smaller histogram is zoomed-in on count depths below 4,000. A threshold is applied here at 1,500 based on the peak detected at around 1,200 counts. (B) Histogram of the number of genes detected per cell. A small noise peak is visible at approx. 400 genes. These cells are filtered out using the depicted threshold (red line) at 700 genes. (C) Count depth distribution from high to low count depths. This visualization is related to the log–log plot shown in Cell Ranger outputs that is used to filter out empty droplets. It shows an "elbow" where count depths start to decrease rapidly around 1,500 counts. (D) Number of genes versus the count depth coloured by the fraction of mitochondrial reads. Mitochondrial read fractions are only high in particularly low count cells with few detected genes. These cells are filtered out by our count and gene number thresholds. Jointly visualizing the count and gene thresholds shows the joint filtering effect, indicating that a lower gene threshold may have sufficed. [Correction added on 5 July 2019, after first online publication: the *x*-axis label in panel B was corrected from "Count depth" to "Number of genes".]

of interest and leaving some leeway for dropout effects. For example, filtering out genes expressed in fewer than 20 cells may make it difficult to detect cell clusters with fewer than 20 cells. For datasets with high dropout rates, this threshold may also complicate the detection of larger clusters. The choice of threshold should scale with the number of cells in the dataset and the intended downstream analysis.

Further QC can be performed on the count data directly. *Ambient gene expression* refers to counts that do not originate from a barcoded cell, but from other lysed cells whose mRNA contaminated the cell suspension prior to library construction. These added ambient counts can distort downstream analysis such as marker gene identification or other differential expression tests especially when levels vary between samples. It is possible to correct for these effects in droplet-based scRNA-seq datasets due to the large numbers of empty droplets, which can be used to model ambient RNA expression profiles. The recently developed SoupX (preprint: Young & Behjati, 2018) uses this approach to directly correct the count data. Pragmatic approaches that ignore strongly ambient genes in downstream analysis have also been used to tackle this problem (Angelidis *et al*, 2019).

Quality control is performed to ensure that the data quality is sufficient for downstream analysis. As "sufficient data quality" cannot be determined *a priori*, it is judged based on downstream analysis performance (e.g., cluster annotation). Thus, it may be necessary to revisit quality control decisions multiple times when analysing the data. Often it is beneficial to start with permissive QC thresholds and investigate the effects of these thresholds before going back to perform more stringent QC. This approach is particularly relevant

 © 2019 The Authors

for datasets containing heterogeneous cell populations where cell types or states may be misinterpreted as low-quality outlier cells. In low-quality datasets, stringent QC thresholds may be necessary. The quality of a dataset can be determined by experimental QC metrics (see Appendix Supplementary Text S2). In this iterative QC optimization, one should be aware of *data peeking*. QC thresholds should not be adapted to improve the outcome of a statistical test. Instead, QC can be evaluated from the distribution of QC covariates in dataset visualizations and clustering.

Pitfalls & recommendations:

- Perform QC by finding outlier peaks in the number of genes, the count depth and the fraction of mitochondrial reads. Consider these covariates jointly instead of separately.
- Be as permissive of QC thresholding as possible, and revisit QC if downstream clustering cannot be interpreted.
- If the distribution of QC covariates differ between samples, QC thresholds should be determined separately for each sample to account for sample quality differences as in Plasschaert *et al* (2018).

Normalization

Each count in a count matrix represents the successful capture, reverse transcription and sequencing of a molecule of cellular mRNA (Box 1). Count depths for identical cells can differ due to the variability inherent in each of these steps. Thus, when gene expression is compared between cells based on count data, any difference may have arisen solely due to sampling effects. Normalization addresses this issue by e.g. scaling count data to obtain correct relative gene expression abundances between cells.

Many normalization methods exist for bulk gene expression (preprint: Pachter, 2011; Dillies *et al*, 2013). While some of these methods have been applied to scRNA-seq analysis, sources of variation specific to single-cell data such as *technical dropouts* (zero counts due to sampling) have prompted the development of scRNA-seq-specific normalization methods (Lun *et al*, 2016a; Vallejos *et al*, 2017).

The most commonly used normalization protocol is count depth scaling, also referred to as "counts per million" or CPM normalization. This protocol comes from bulk expression analysis and normalizes count data using a so-called size factor proportional to the count depth per cell. Variations of this method scale the size factors with different factors of 10, or by the median count depth per cell in the dataset. CPM normalization assumes that all cells in the dataset initially contained an equal number of mRNA molecules and count depth differences arise only due to sampling. This assumption is shared with the downsampling protocol, which is the process of randomly sampling reads or counts from the data to leave all cells with a prespecified number of counts or fewer. While downsampling throws away data, it also increases technical dropout rates which CPM and other global scaling normalization methods do not. Thus, downsampling can deliver a more realistic representation of what cellular expression profiles would look like at similar count depths.

As single-cell datasets typically consist of heterogeneous cell populations with varying sizes and molecule counts, more complex normalization methods are usually appropriate. For example, Weinreb *et al* (2018) used a simple extension of CPM that excludes genes that account for at least 5% of the total counts in any cell, when calculating their size factors. This approach allows for molecular count variability in few highly expressed genes. More cellular heterogeneity is permitted by Scran's pooling-based size factor estimation method (Lun et al, 2016a). Here, size factors are estimated based on a linear regression over genes, after cells are pooled to avoid technical dropout effects. This method limits variability to fewer than 50% of genes being differentially expressed between cells, and is consistently a top-performing normalization method in independent comparisons. Scran has been shown to perform better than other tested normalization methods for batch correction (Buttner *et al*, 2019) and differential expression analysis (preprint: Vieth *et al*, 2019). The method was also shown to give robust size factor estimates in a small-scale comparison from the original authors (Vallejos *et al*, 2017).

CPM, high-count filtering CPM, and scran use linear, global scaling to normalize count data. Non-linear normalization methods, which can account for more complex unwanted variation, also exist (Cole *et al*, 2019). Many such methods involve the parametric modelling of count data. For example, Mayer *et al* (2018) fit a negative binomial model to count data, using technical covariates such as the read depth and the number of counts per gene to fit the model parameters. The residuals of the model fit serve as a normalized quantification of gene expression. Such an approach can combine technical and biological data correction (e.g. batch correction or correction for cell cycle effects) with count depth normalization. Non-linear normalization methods have been shown to outperform global scaling methods especially in situations with strong batch effects (Cole *et al*, 2019). Thus, non-linear normalization methods are particularly relevant for plate-based scRNA-seq data, which tend to have batch effects between plates. Furthermore, plate-based data can exhibit larger variations in count depths per cell than droplet-based data (Svensson *et al*, 2017). While non-linear normalization methods, or alternative approaches such as downsampling, appear better suited to these conditions, comparative studies are needed to confirm this hypothesis. In this tutorial, we prefer to separate the normalization and data correction (batch correction, noise correction, etc.) steps to emphasize different processing stages of the data (see "Stages of pre-processed data" section). Thus, we focus on global scaling normalization methods.

We cannot expect that a single normalization method is appropriate for all types of scRNA-seq data. For example, Vieth *et al* (2017) showed that read and count data are best fit by different models. Indeed Cole *et al* (2019) find that different normalization methods perform optimally for different datasets and argue that their *scone* tool should be used to select the appropriate normalization method for a specific dataset. Furthermore, scRNA-seq techniques can be divided into full-length and 3′ enrichment methods (Svensson *et al*, 2017; Ziegenhain *et al*, 2017). Data from full-length protocols may benefit from normalization methods that take into account gene length (e.g. Patel *et al*, 2014; Kowalczyk *et al*, 2015; Soneson & Robinson, 2018), while 3′ enrichment data do not. A commonly used normalization method for full-length scRNA-seq data is TPM normalization (Li *et al*, 2009), which comes from bulk RNA-seq analysis.

In the same way that cellular count data can be normalized to make them comparable between cells, gene counts can be scaled to improve comparisons between genes. Gene normalization constitutes scaling gene counts to have zero mean and unit variance (z scores). This scaling has the effect that all genes are

© 2019 The Authors

weighted equally for downstream analysis. There is currently no consensus on whether or not to perform normalization over genes. While the popular Seurat tutorials (Butler *et al*, 2018) generally apply gene scaling, the authors of the Slingshot method opt against scaling over genes in their tutorial (Street *et al*, 2018). The preference between the two choices revolves around whether all genes should be weighted equally for downstream analysis, or whether the magnitude of expression of a gene is an informative proxy for the importance of the gene. In order to retain as much biological information as possible from the data, we opt to refrain from scaling over genes in this tutorial.

After normalization, data matrices are typically log(x + 1)-transformed. This transformation has three important effects. Firstly, distances between log-transformed expression values represent log fold changes, which are the canonical way to measure changes in expression. Secondly, log transformation mitigates (but does not remove) the mean–variance relationship in single-cell data (Brennecke *et al*, 2013). Finally, log transformation reduces the skewness of the data to approximate the assumption of many downstream analysis tools that the data are normally distributed. While scRNA-seq data are not in fact log-normally distributed (Vieth *et al*, 2017), these three effects make the log transformation a crude, but useful tool. This usefulness is highlighted by downstream applications for differential expression testing (Finak *et al*, 2015; Ritchie *et al*, 2015) or batch correction (Johnson *et al*, 2006; Buttner *et al*, 2019) that use log transformation for these purposes. It should however be noted that log transformation of normalized data can introduce spurious differential expression effects into the data (preprint: Lun, 2018). This effect is particularly pronounced when normalization size factor distributions differ strongly between tested groups.

Pitfalls & recommendations:

- We recommend scran for normalization of non-full-length datasets. An alternative is to evaluate normalization approaches via *scone* especially for plate-based datasets. Full-length scRNA-seq protocols can be corrected for gene length using bulk methods.
- There is no consensus on scaling genes to 0 mean and unit variance. We prefer not to scale gene expression.
- Normalized data should be log(x+1)-transformed for use with downstream analysis methods that assume data are normally distributed.

Data correction and integration

Normalization as described above attempts to remove the effects of count sampling. However, normalized data may still contain unwanted variability. Data correction targets further technical and biological covariates such as batch, dropout, or cell cycle effects. These covariates are not always corrected for. Instead, the decision of which covariates to consider will depend on the intended downstream analysis. We propose to consider correction for biological and technical covariates separately as these are used for different purposes and present unique challenges.

Regressing out biological effects

While correcting for technical covariates may be crucial to uncovering the underlying biological signal, correction for biological covariates serves to single out particular biological signals of interest. The most common biological data correction is to remove the effects of the cell cycle on the transcriptome. This data correction can be performed by a simple linear regression against a cell cycle score as implemented in the Scanpy and Seurat platforms (Butler *et al*, 2018; Wolf *et al*, 2018) or in specialized packages with more complex mixture models such as scLVM (Buettner *et al*, 2015) or f-scLVM (Buettner *et al*, 2017). Lists of marker genes to compute cell cycle scores are obtained from the literature (Macosko *et al*, 2015). These methods can also be used to regress out other known biological effects such as mitochondrial gene expression, which is interpreted as an indication of cell stress.

Several aspects should be considered prior to correcting data for biological effects. Firstly, correcting for biological covariates is not always helpful to interpret scRNA-seq data. While removing cell cycle effects can improve the inference of developmental trajectories (Buettner *et al*, 2015; Vento-Tormo *et al*, 2018), cell cycle signals can also be informative of the biology. For example, proliferating cell populations can be identified based on cell cycle scores (see case study on project github). Also, biological signals must be understood in context. Given that biological processes occur within the same organism, there exist dependencies between these processes. Thus, correcting for one process may unintentionally mask the signal of another. Finally, it has been argued that variation in cell size accounts for the transcriptomic effect generally attributed to the cell cycle (McDavid *et al*, 2016). Thus, correcting for cell size via normalization, or dedicated tools such as cgCorrect (Blasi *et al*, 2017), also partially corrects for cell cycle effects in scRNA-seq data.

Regressing out technical effects

The variants of regression models used to regress out biological covariates can also be applied to technical covariates. The most prominent technical covariates in single-cell data are count depth and batch. Although normalization scales count data to render gene counts comparable between cells, a count depth effect often remains in the data. This count depth effect can be both a biological and a technical artefact. For example, cells may differ in size and therefore in mRNA molecule counts. Yet, technical count effects may remain after normalization as no scaling method can infer the expression values of genes that were not detected due to poor sampling. Regressing out count depth effects can improve the performance of trajectory inference algorithms, which rely on finding transitions between cells (see case study on project github). When correcting for multiple covariates (e.g. cell cycle and count depth), the regression should be performed over all covariates in a single step to account for dependence between covariates.

An alternative to regression-based strategies for removing count effects is to use a more rigorous normalization procedure such as downsampling or non-linear normalization methods (see "Normalization" section). These approaches may be particularly relevant for plate-based scRNA-seq datasets where the larger variation of count depths per cell can mask the heterogeneity between cells.

Batch effects and data integration

Batch effects can occur when cells are handled in distinct groups. These groups can consist of cells on different chips, cells in different sequencing lanes or cells harvested at different time

 © 2019 The Authors

points. The differing environments experienced by the cells can have an effect on the measurement of the transcriptome or on the transcriptome itself. The resulting effects exist on multiple levels: between groups of cells in an experiment, between experiments performed in the same laboratory or between datasets from different laboratories. Here, we distinguish between the first and the last two scenarios. Correcting for batch effects between samples or cells in the same experiment is the classical scenario known as *batch correction* from bulk RNA-seq. We distinguish this from the integration of data from multiple experiments, which we call *data integration*. While batch effects are typically corrected using linear methods, non-linear approaches are used for data integration.

A recent comparison of classical batch correction methods has revealed that ComBat (Johnson *et al*, 2006) performs well also for single-cell experiments of low-to-medium complexity (Buttner *et al*, 2019). ComBat consists of a linear model of gene expression where the batch contribution is taken into account both in the mean and the variance of the data (Fig 3). Irrespective of computational methods, the best method of batch correction is pre-empting the effect and avoiding it altogether by clever experimental design (Hicks *et al*, 2017). Batch effects can be avoided by pooling cells across experimental conditions and samples. Using strategies such as cell tagging (preprint: Gehring *et al*, 2018), or via genetic variation (Kang *et al*, 2018), it is possible to demultiplex cells that were pooled in the experiment.

In comparison with batch correction, the additional challenge that data integration methods face revolves around compositional differences between datasets. When estimating batch effects, ComBat uses all cells in a batch to fit batch parameters. This approach will confound the batch effect with biological differences between cell types or states that are not shared among datasets. Data integration methods such as Canonical Correlation Analysis (CCA; Butler *et al*, 2018), Mutual Nearest Neighbours (MNN; Haghverdi *et al*, 2018), Scanorama (preprint: Hie *et al*, 2018), RISC (preprint: Liu *et al*, 2018), scGen (preprint: Lotfollahi *et al*, 2018), LIGER (preprint: Welch *et al*, 2018), BBKNN (preprint: Park *et al*, 2018), and Harmony (preprint: Korsunsky *et al*, 2018) have been developed to overcome this issue. While data integration methods can also be applied to simple batch correction problems, we recommend to be wary of over-correction given the increased degrees of freedom of non-linear data integration approaches. For example, MNN was shown to be outperformed by ComBat in the simpler batch correction setting (Buttner *et al*, 2019). Further comparison studies between data integration and batch correction methods are needed to assess how generally these methods can be applied.

Expression recovery

A further type of technical data correction is *expression recovery* (also *denoising* or *imputation*). Measurements of single-cell transcriptomes contain various sources of noise (Grün *et al*, 2014; Kharchenko *et al*, 2014; Hicks *et al*, 2017). A particularly prominent aspect of this noise is dropout. Inferring dropout events, replacing these zeros with appropriate expression values, and reducing the noise in the dataset have been the target of several

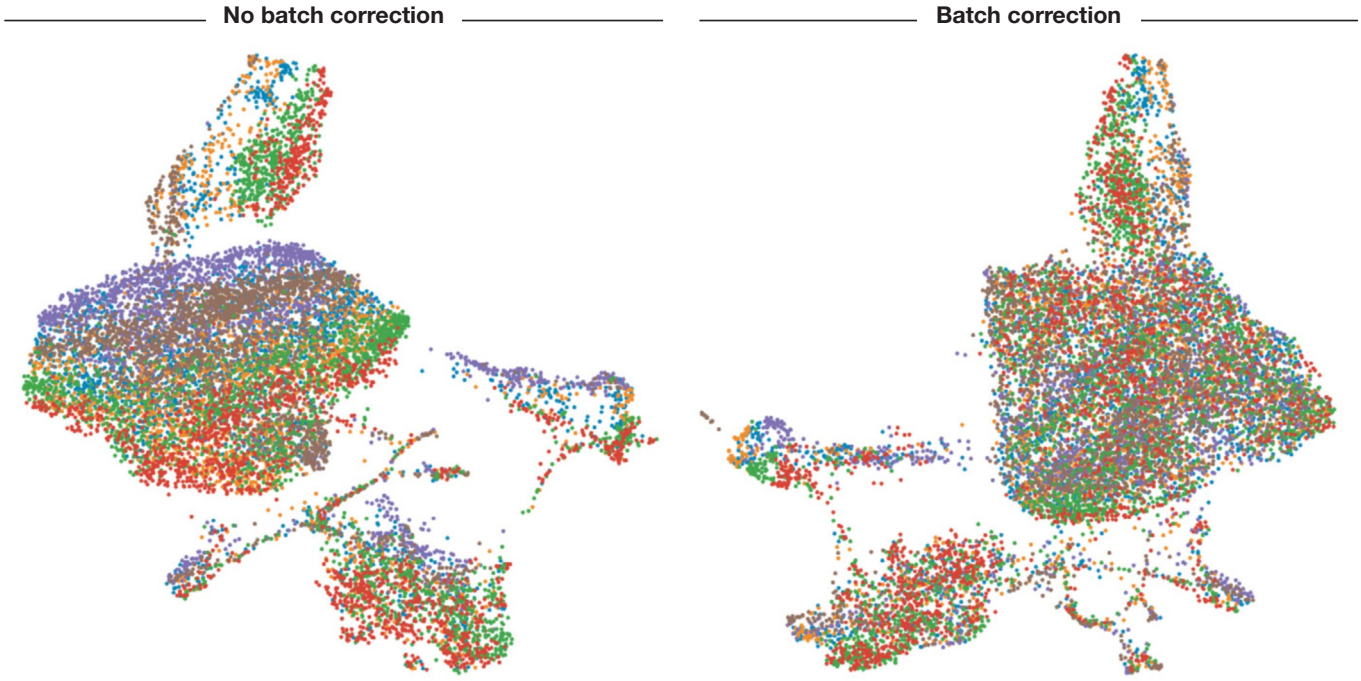


Figure 3. UMAP visualization before and after batch correction.

Cells are coloured by sample of origin. Separation of batches is clearly visible before batch correction and less visible afterwards. Batch correction was performed using ComBat on mouse intestinal epithelium data from Haber *et al* (2017).

© 2019 The Authors

recent tools (MAGIC: van Dijk *et al*, 2018; DCA: Eraslan *et al*, 2019; scVI: Lopez *et al*, 2018; SAVER: Huang *et al*, 2018; scImpute: Li & Li, 2018). Performing expression recovery has been shown to improve the estimation of gene–gene correlations (van Dijk *et al*, 2018; Eraslan *et al*, 2019). Furthermore, this step can be integrated with normalization, batch correction and other downstream analysis as implemented in the scVI tool (Lopez *et al*, 2018). While most data correction methods take normalized data as input, some expression recovery methods are based on expected negative binomial noise distributions and therefore run on raw count data. When applying expression recovery, one should take into consideration that no method is perfect. Thus, any method may over- or under-correct noise in the data. Indeed, false correlation signals have been reported as a result of expression recovery (Andrews & Hemberg, 2018). Given the difficulty of assessing successful expression recovery in a practical application, this scenario represents a challenge to the user pondering whether or not to denoise their data. Furthermore, scalability to large datasets is still an issue for currently available expression recovery methods. There is currently no consensus on how denoised data should be used in the light of these considerations (see "Stages of processed data" section). A prudent approach would be to use expression recovery only for visual display of data rather than to generate hypotheses during exploratory data analysis. Thorough experimental validation is particularly important here.

Pitfalls & recommendations:

- Regress out biological covariates only for trajectory inference and if other biological processes of interest are not masked by the regressed out biological covariate.
- Regress out technical and biological covariates jointly rather than serially.
- Plate-based dataset pre-processing may require regressing out counts, normalization via non-linear normalization methods or downsampling.
- We recommend performing batch correction via ComBat when cell type and state compositions between batches are consistent
- Data integration and batch correction should be performed by different methods. Data integration tools may over-correct simple batch effects.
- Users should be cautious of signals found only after expression recovery. Exploratory analysis may be best performed without this step.

Feature selection, dimensionality reduction and visualization

A human single-cell RNA-seq dataset can contain expression values for up to 25,000 genes. Many of these genes will not be informative for a given scRNA-seq dataset, and many genes will mostly contain zero counts. Even after filtering out these zero count genes in the QC step, the feature space for a single-cell dataset can have over 15,000 dimensions. To ease the computational burden on downstream analysis tools, reduce the noise in the data, and to visualize the data, one can use several approaches to reduce the dimensionality of the dataset.

Feature selection

The first step of reducing the dimensionality of scRNA-seq datasets commonly is feature selection. In this step, the dataset is filtered to keep only genes that are "informative" of the variability in the data. Thus, highly variable genes (HVGs) are often used (Brennecke *et al*, 2013). Depending on the task and the complexity of the dataset, typically between 1,000 and 5,000 HVGs are selected for downstream analysis (see Fig EV1 and Dataset EV1). Preliminary results from Klein *et al* (2015) suggest that downstream analysis is robust to the exact choice of the number of HVGs. While varying the number of HVGs between 200 and 2,400, the authors reported similar low-dimensional representations in the PCA space. Based on this result, we prefer to err on the side of higher numbers of HVGs.

A simple, yet popular, method of selecting HVGs is implemented in both Scanpy and Seurat. Here, genes are binned by their mean expression, and the genes with the highest variance-to-mean ratio are selected as HVGs in each bin. There exist different flavours of this algorithm that expect either count data (Seurat) or log-transformed data (Cell Ranger). Optimally, HVGs should be selected after technical data correction to avoid selecting genes that are highly variable only due to, e.g., batch effects. Other methods for HVG selection are reviewed in Yip *et al* (2018).

Dimensionality reduction

After feature selection, the dimensions of single-cell expression matrices can be further reduced by dedicated dimensionality reduction algorithms. These algorithms embed the expression matrix into a low-dimensional space, which is designed to capture the underlying structure in the data in as few dimensions as possible. This approach works as single-cell RNA-seq data are inherently low-dimensional (Heimberg *et al*, 2016). In other words, the biological manifold on which cellular expression profiles lie can be sufficiently described by far fewer dimensions than the number of genes. Dimensionality reduction aims to find these dimensions.

There are two main objectives of dimensionality reduction methods: *visualization* and *summarization*. Visualization is the attempt to optimally describe the dataset in two or three dimensions. These reduced dimensions are used as coordinates on a scatter plot to obtain a visual representation of the data. Summarization does not prescribe the number of output components. Instead, higher components become less important for describing the variability present in the data. Summarization techniques can be used to reduce the data to its essential components by finding the inherent dimensionality of the data, and are thus helpful for downstream analysis. While a 2-dimensional visualization output should not be used to summarize a dataset, a summarization method can be used to visualize the data using the leading reduced components. However, a dedicated visualization technique will typically provide a better representation of the variability.

Reduced dimensions are generated through linear or non-linear combinations of feature space dimensions (gene expression vectors). Especially in the non-linear case, the interpretability of the reduced dimensions is sacrificed in this process. An example application of some commonly used dimensionality reduction methods is shown in Fig 4. With a growing list of methods to choose from, it is out of the scope of this tutorial to review these methods in detail. Rather, we briefly outline the practical considerations that may aid users in choosing between common dimensionality reduction methods. A more detailed review of dimensionality reduction for single-cell analysis can be found in Moon *et al* (2018).

 © 2019 The Authors

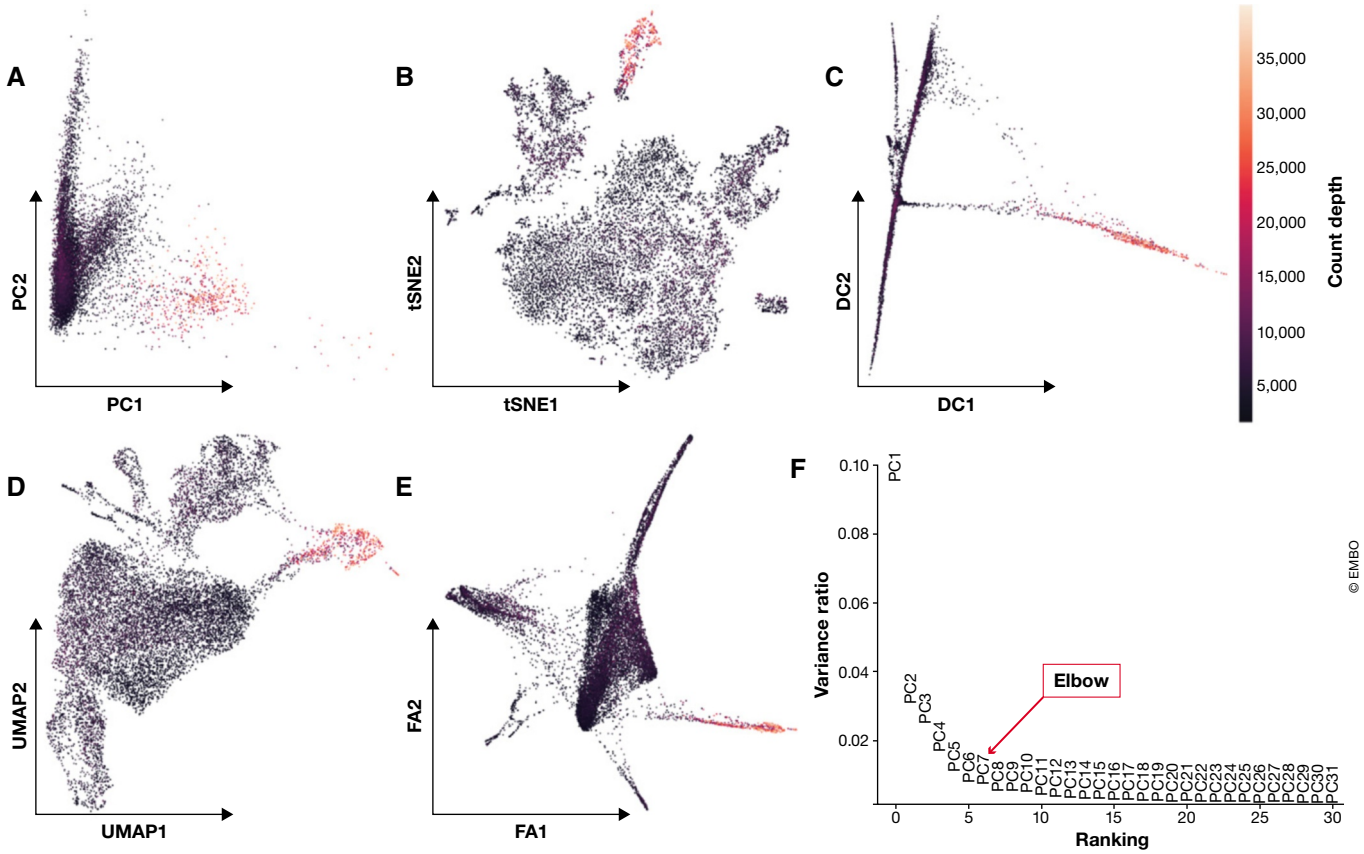


Figure 4. Common visualization methods for scRNA-seq data.

Mouse intestinal epithelium regions data from Haber *et al* (2017) visualized on the first two components for: (A) PCA, (B) t-SNE, (C) diffusion maps, (D) UMAP and (E) A force-directed graph layout via ForceAtlas2. Cells are coloured by count depth. (F) Variance explained by the first 31 principal components (PCs). The "elbow" of this plot, which is used to select relevant PCs to analyse the dataset, lies between PCs 5 and 7.

Two popular dimensionality reduction techniques that are principally summarization methods are principal component analysis (PCA; Pearson, 1901) and diffusion maps (Coifman *et al*, 2005) which were popularized for single-cell analysis by Haghverdi *et al* (2015). Principal component analysis is a linear approach that generates reduced dimensions by maximizing the captured residual variance in each further dimension. While PCA does not capture the structure of the data in few dimensions as well as non-linear methods, it is the basis of many currently available analysis tools for clustering or trajectory inference. Indeed, PCA is commonly used as a pre-processing step for non-linear dimensionality reduction methods. Typically, PCA summarizes a dataset via its top N principal components, where N can be determined by "elbow" heuristics (see Fig 4F) or the permutation-test-based jackstraw method (Chung & Storey, 2015; Macosko *et al*, 2015). The simple linearity of PCA has the advantage that distances in reduced dimensional space have a consistent interpretation in all regions of this space. Thus, we can correlate quantities of interest with principal components to assess their importance. For example, principal components can be projected onto technical nuisance covariates to investigate the performance of QC, data correction and normalization steps (Buttner *et al*, 2019), or show the importance of genes in the dataset (Chung & Storey, 2015). Diffusion maps are a non-linear data summarization technique. As diffusion components emphasize transitions in the data, they are principally used when continuous processes such as differentiation are of interest. Typically, each diffusion component (i.e. diffusion map dimension) highlights the heterogeneity of a different cell population.

Visualization

For visualization purposes, it is standard practice to use non-linear dimensionality reduction methods (Fig 4). The most common dimensionality reduction method for scRNA-seq visualization is the t-distributed stochastic neighbour embedding (t-SNE; van der Maaten & Hinton, 2008). t-SNE dimensions focus on capturing local similarity at the expense of global structure. Thus, these visualizations may exaggerate differences between cell populations and overlook potential connections between these populations. A further difficulty is the choice of its perplexity parameter, as t-SNE graphs may show strongly different numbers of clusters depending on its value (Wattenberg *et al*, 2016). Common alternatives to t-SNE are the Uniform Approximation and Projection method (UMAP; preprint: McInnes & Healy, 2018) or graph-based tools such as SPRING (Weinreb *et al*, 2018). UMAP and SPRING's force-directed layout algorithm ForceAtlas2 arguably represent the best approximation of the underlying topology (Wolf *et al*, 2019, Supplemental

© 2019 The Authors

Table 1. Stages of data processing and appropriate downstream applications.

Pre-processing layer	Stage of data processing	Appropriate applications
Measured	1) Raw	Statistical testing (Differential expression: marker genes, genes over condition, genes over time)
	2) Normalized (+ log transformed)	
Corrected	3.1) Corrected (technical correction)	Visual comparison of data (plotting)
	3.2) Corrected (biological correction)	Pre-processing for trajectory inference
Reduced	4) Feature selected	Visualization, trajectory inference
	5) Dimensionality reduced (summarized)	Visualization, clustering, KNN graph inference, trajectory inference

Note 4). What sets UMAP apart in this comparison is its speed and ability to scale to large numbers of cells (Becht *et al*, 2018). Thus, in the absence of particular biological questions, we regard UMAP as best practice for exploratory data visualization. Moreover, UMAP can also summarize data in more than two dimensions. While we are not aware of any applications of UMAP for data summarization, it may prove a suitable alternative to PCA.

An alternative to classical visualization on the cell level is partition-based graph abstraction (PAGA; Wolf *et al*, 2019). This tool has been shown to adequately approximate the topology of the data while coarse-graining the visualization using clusters. In combination with any of the above visualization methods, PAGA produces coarse-grained visualizations, which can simplify the interpretation of single-cell data especially with large numbers of cells.

Pitfalls & recommendations:

- We recommend selecting between 1,000 and 5,000 highly variable genes depending on dataset complexity.
- Feature selection methods that use gene expression means and variances cannot be used when gene expression values have been normalized to zero mean and unit variance, or when residuals from model fitting are used as normalized expression values. Thus, one must consider what pre-processing to perform before selecting HVGs.
- Dimensionality reduction methods should be considered separately for summarization and visualization.
- We recommend UMAP for exploratory visualization; PCA for general purpose summarization; and diffusion maps as an alternative to PCA for trajectory inference summarization.
- PAGA with UMAP is a suitable alternative to visualize particularly complex datasets.

Stages of pre-processed data

While we have outlined common pre-processing steps in scRNA-seq as a sequential pipeline above, downstream analyses often prefer to take different levels of pre-processed data, and it is recommended to adapt pre-processing depending on the downstream application. To clarify this situation to a new user, we delineated pre-processing into five stages of data processing: (i) raw data, (ii) normalized data, (iii) corrected data, (iv) feature-selected data, and (v) dimensionality-reduced data. These stages of data processing are grouped into three pre-processing layers: measured data, corrected data, and reduced data. Cell and gene QC should always be performed and is therefore omitted from this characterization. While the order of the layers represent the typical workflow in scRNA-seq analysis, it is also possible to skip layers or have slight alterations in the order of processing stages. For example, data correction may not be necessary for single batch datasets. In Table 1, we summarize the appropriate downstream applications for each layer of pre-processed data.

The stages of pre-processing in Table 1 are divided into three groups: measured data, corrected data and reduced data. We define *measured data* as raw data and processed data that retain the structure of zeros. By scaling count data with a cell-specific factor, global scaling normalization methods retain zero expression values even after $\log(x+1)$-transformation. In contrast, correcting data for unwanted variability replaces zero expression values. The *corrected data* layer represents the "cleanest" version of the data, which is the closest approximation of the underlying biological signal. We call the final pre-processing layer *reduced data*. This data layer emphasizes dominant aspects of the data, which can be described using a reduced set of features.

The aforementioned characteristics determine the suitability of the pre-processed data for particular downstream applications. As the final pre-processing stage, reduced data would be the natural candidate for a broadly applicable data layer. However, testing for differential expression is only biologically interpretable in gene space, which is not (fully) represented in reduced data. The strength of reduced data lies in the summarization of the biology and the reduction of noise, which can mask biological signals. Thus, reduced data are used for exploratory methods that require data summaries (visualization, neighbourhood graph inference, clustering) and for computationally complex downstream analysis tools (trajectory inference). Indeed, many trajectory inference methods incorporate dimensionality reduction in the tools themselves.

The expression profiles of individual genes can only be compared in gene space, which is captured in measured and corrected data. Comparison of expression profiles can be performed visually and statistically. We argue that visual and statistical comparison should be performed on different data layers. For visual inspection of gene expression, corrected data are most appropriate. Should raw data be presented for visual comparison, the user is required to inherently understand the biases in the data in order to interpret the results. Corrected data facilitate this interpretation. However, one should consider corrected data for technical and biological covariates separately here. While correction for biological covariates may increase the strength of a particular biological signal, it produces a less

 © 2019 The Authors

accurate representation of the underlying biology and will mask other signals that may be relevant. Thus, biologically corrected data are appropriate mainly for analysis tools that focus on particular biological processes such as trajectory inference methods.

Statistical comparison of gene expression is most appropriate on the measured data layer. No perfect data correction method exists for denoising, batch correction or correcting for other sources of variation. Thus, data correction methods inevitably over- or under-correct the data and therefore alter the variance of at least some gene expression profiles in an unintended way. Statistical tests of gene expression rely on assessing the background variance as a null model for noise in the data. As data correction tends to reduce background variation (Fig EV2), genes whose background variation is over-corrected by data correction methods will be more likely assessed as significantly differentially expressed. Furthermore, certain data correction methods (e.g. ComBat) interpret expression signals that do not conform to the experimental design as noise, which is subsequently removed from the data. In addition to an underestimation of the noise, this optimization of the experimental design signal can lead to an overestimation of the effect size. In the light of these considerations, using measured data as input, as opposed to corrected data, constitutes a more conservative approach to differential testing. With measured data, technical covariates can and should be taken into consideration in the differential testing model.

The above view is supported by a recent comparison of scRNA-seq differential expression methods, which uses only raw and normalized data as input (Soneson & Robinson, 2018). The normalized data used in this study revolve only around global scaling methods. However, many currently available non-linear normalization methods blur the line between normalization and data correction (see "Normalization" section). Such normalized data may no longer be appropriate as input for differential testing.

Pitfalls & recommendations:

- Use measured data for statistical testing, corrected data for visual comparison of data and reduced data for other downstream analysis based on finding the underlying biological data manifold.

Downstream analysis

After pre-processing, methods that we call *downstream analysis* are used to extract biological insights and describe the underlying biological system. These descriptions are obtained by fitting interpretable models to the data. Examples of such models are groups of cells with similar gene expression profiles representing cell-type clusters; small changes in gene expression between similar cells denoting continuous (differentiation) trajectories; or genes with correlated expression profiles indicating co-regulation.

Downstream analysis can be divided into cell- and gene-level approaches as shown in Fig 5. Cell-level analysis typically focuses on the description of two structures: clusters and trajectories. These structures can in turn be analysed on the cell and the gene level leading to cluster analysis and trajectory analysis methods. Broadly, cluster analysis methods attempt to explain the heterogeneity in the data based on a categorization of cells into groups. In contrast, in trajectory analysis the data are regarded as a snapshot of a dynamic process. Trajectory analysis methods investigate this underlying process.

Here, we describe cell- and gene-level cluster and trajectory analysis tools before detailing gene-level analyses that are performed independently of these cellular structures.

Cluster analysis

Clustering

Organizing cells into clusters is typically the first intermediate result of any single-cell analysis. Clusters allow us to infer the identity of member cells. Clusters are obtained by grouping cells based on the similarity of their gene expression profiles. Expression profile similarity is determined via distance metrics, which often take dimensionality-reduced representations as input. A common example of similarity scoring is Euclidean distances that are calculated on the PC-reduced expression space. Two approaches exist to generate cell clusters from these similarity scores: clustering algorithms and community detection methods.

Clustering is a classical unsupervised machine learning problem, based directly on a distance matrix. Cells are assigned to clusters by minimizing intracluster distances or finding dense regions in the reduced expression space. The popular *k*-means clustering algorithm divides cells into *k* clusters by determining cluster centroids and assigning cells to the nearest cluster centroid. Centroid positions are iteratively optimized (MacQueen, 1967). This approach requires an input of the number of clusters expected, which is usually unknown and must be calibrated heuristically. Applications of k-means to single-cell data vary in the distance metrics used. Alternatives to standard Euclidean distances include cosine similarity (Haghverdi *et al*, 2018), correlation-based distance metrics (Kim *et al*, 2018) or the SIMLR method, which learns a distance metric for each dataset using Gaussian kernels (Wang *et al*, 2017). A recent comparison has suggested that correlation-based distances may outperform other distance metrics when used with k-means or as the basis for Gaussian kernels (Kim *et al*, 2018).

Community detection methods are graph-partitioning algorithms and thus rely on a graph representation of single-cell data. This graph representation is obtained using a *K*-Nearest Neighbour approach (KNN graph). Cells are represented as nodes in the graph. Each cell is connected to its *K* most similar cells, which are typically obtained using Euclidean distances on the PC-reduced expression space. Depending on the size of the dataset, *K* is commonly set to be between 5 and 100 nearest neighbours. The resulting graph captures the underlying topology of the expression data (Wolf *et al*, 2019). Densely sampled regions of expression space are represented as densely connected regions of the graph. These dense regions are detected using community detection methods. Community detection is often faster than clustering as only neighbouring cell pairs have to be considered as belonging to the same cluster. This approach thus greatly reduces the search space for possible clusters.

After the pioneering PhenoGraph method (Levine *et al*, 2015), the standard approach to clustering single-cell datasets has become

© 2019 The Authors

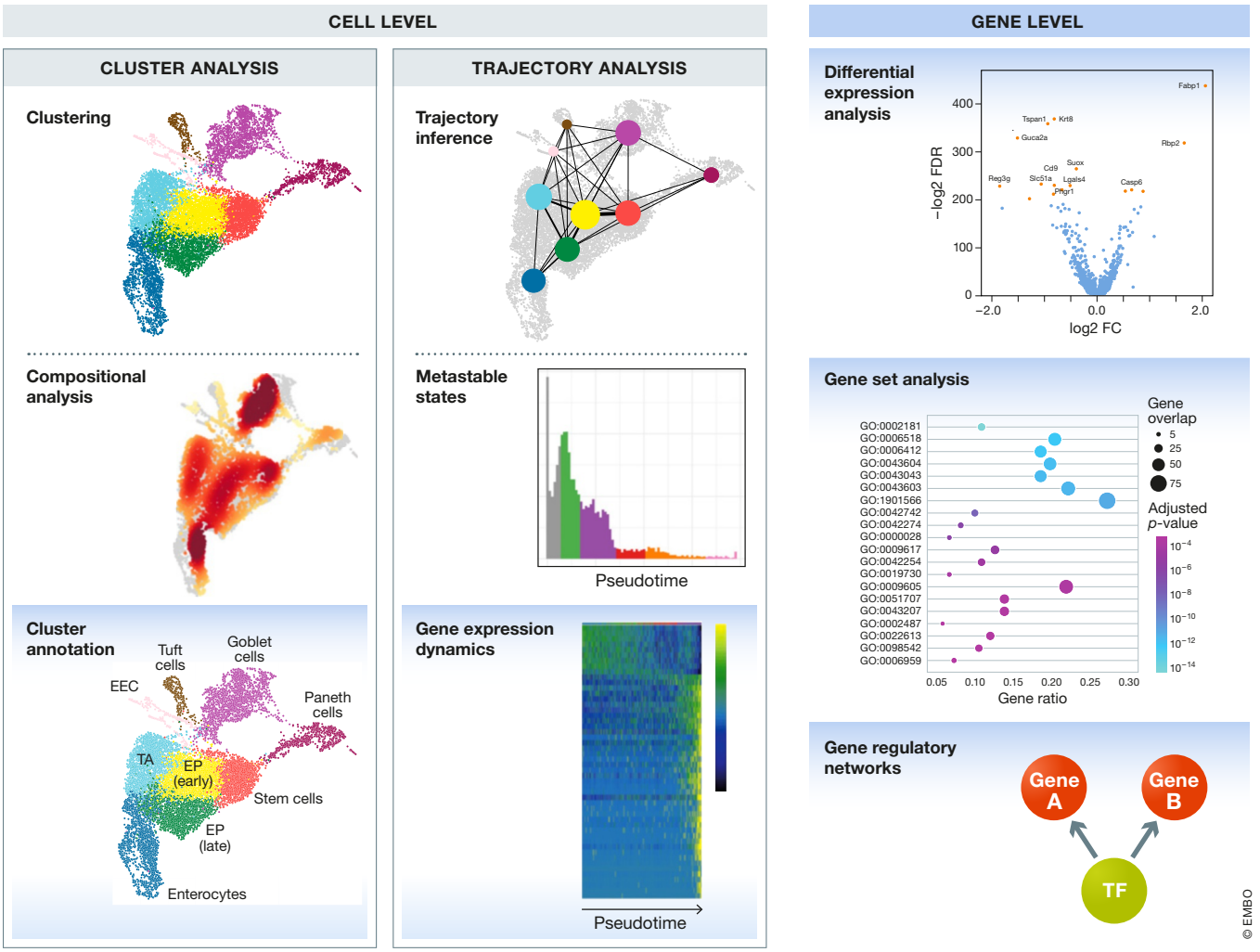


Figure 5. Overview of downstream analysis methods.
Methods are divided into cell- and gene-level analysis. Cell-level analysis approaches are again subdivided into cluster and trajectory analysis branches, which include also gene-level analysis methods. All methods with a blue background are gene-level approaches.

multi-resolution modularity optimization (Newman & Girvan, 2004; Reichardt & Bornholdt, 2006) as implemented in the Louvain algorithm (Blondel *et al*, 2008) on single-cell KNN graphs. This method is the default clustering method implemented in the Scanpy and Seurat single-cell analysis platforms. It has been shown to outperform other clustering methods for single-cell RNA-seq data (Duò *et al*, 2018; Freytag *et al*, 2018), and flow and mass cytometry data (Weber & Robinson, 2016). Conceptually, the Louvain algorithm detects communities as groups of cells that have more links between them than expected from the number of links the cells have in total. The optimized modularity function includes a resolution parameter, which allows the user to determine the scale of the cluster partition. By subsetting the KNN graph, it is also possible to subcluster only particular clusters. Such subclustering can allow the user to identify cell states within cell-type clusters (Wagner *et al*, 2016), but may also lead to patterns that arise only from noise in the data.

Pitfalls & recommendations:

- We recommend clustering by Louvain community detection on a single-cell KNN graph.
- Clustering does not have to be performed at a single resolution. Subclustering particular cell clusters is a valid approach to focus on more detailed substructures in a dataset.

Cluster annotation

On a gene level, clustered data are analysed by finding the gene signatures of each cluster. These so-called *marker genes* characterize the cluster and are used to annotate it with a meaningful biological label. This label represents the identity of cells within the cluster. As any clustering algorithm will produce a partition of the data, the validity of the identified clusters can only be determined by successful annotation of the represented biology.

 © 2019 The Authors

While it may be tempting to assume that the clusters detected in single-cell data represent cell types, there are several axes of variation that determine cellular identity (Wagner *et al*, 2016; Clevers *et al*, 2017). Firstly, it is not always clear what constitutes a cell type. For example, while "T cells" may be a satisfactory label of a cell type to some, others may look for T-cell subtypes within a dataset and distinguish between $CD4^+$ and $CD8^+$ T cells (Wagner *et al*, 2016; Clevers *et al*, 2017). Furthermore, cells of the same cell type in different states may be detected in separate clusters. For the above reasons, it is best to use the term "cell identities" rather than "cell types". Before clustering and annotating clusters, the user must decide which level of annotation detail, and thus which cluster resolution, is of interest.

Identifying and annotating clusters relies on using external sources of information describing the expected expression profiles of individual cell identities. Thanks to recent and ongoing efforts such as the mouse brain atlas (Zeisel *et al*, 2018) or the Human Cell

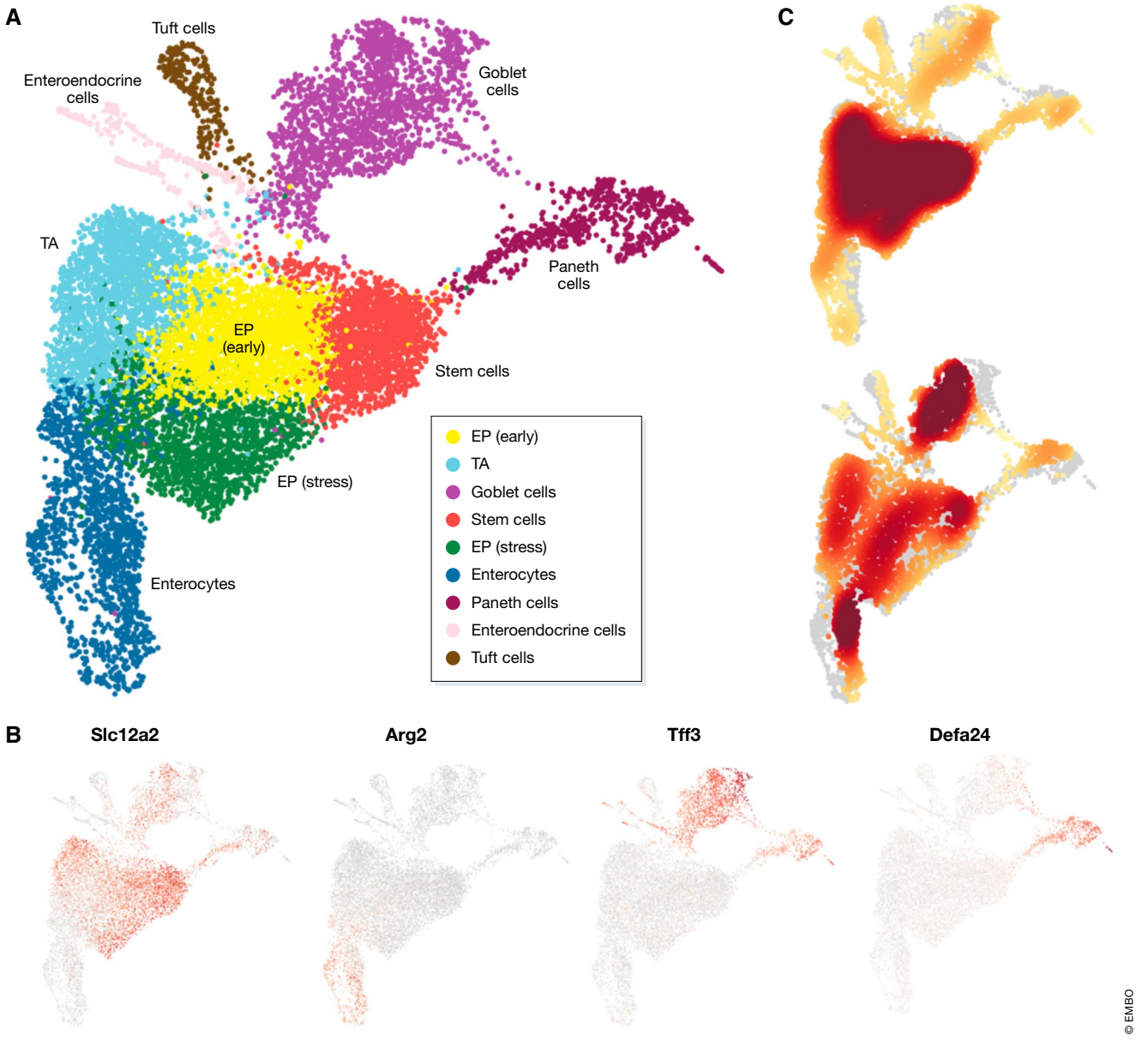


Figure 6. Cluster analysis results of mouse intestinal epithelium dataset from Haber *et al* (2017).

(A) Annotated cell-identity clusters found by Louvain clustering visualized in a UMAP representation. (B) Cell-identity marker expression to identify stem cells (Slc12a2), enterocytes (Arg2), goblet cells (Tff3) and Paneth cells (Defa24). Corrected expression levels are visualized from low expression (grey) to high expression (red). Marker genes may be expressed also in other cell-identity populations as shown for goblet and Paneth cells. (C) Cell-identity composition heat maps of proximal (upper) and distal (lower) intestinal epithelium regions. High relative cell density is shown as dark red.

© 2019 The Authors

Atlas (Regev *et al*, 2017), reference databases are increasingly becoming available. These databases greatly facilitate cell identity annotation. In the absence of a relevant reference database, cell identities can be annotated by comparing data-derived marker genes with marker genes from the literature (see case study on project github) or by directly visualizing the expression values of literature-derived marker genes (Fig 6B). It should be noted that the latter method constrains the user to the classical understanding of cell types derived from bulk expression studies, rather than cell identities. Furthermore, it has been shown that commonly used cell surface markers are limited in their ability to define cell identities (Tabula Muris Consortium *et al*, 2018).

There are two ways to use reference database information to annotate clusters: using data-derived marker genes or using full gene expression profiles. Marker gene sets can be found by applying differential expression (DE) testing between two groups: the cells in one cluster and all other cells in the dataset (see "Differential expression testing"). Typically, we focus on genes that are up-regulated in the cluster of interest. As marker genes are expected to have strong differential expression effects, simple statistical tests such as the Wilcoxon rank-sum test or the *t*-test are often used to rank genes by their difference in expression between these two groups. The top-ranked genes from the respective test statistic are regarded as marker genes. Clusters can be annotated by comparing marker genes from the dataset and marker genes from reference dataset via enrichment tests, the Jaccard index or other overlap statistics. Reference webtools such as www.mousebrain.org (Zeisel *et al*, 2018) or http://dropviz.org/ (Saunders *et al*, 2018) allow users to visualize the expression of dataset marker genes in the reference dataset to facilitate cell-identity annotation.

Two aspects should be noted when detecting marker genes. Firstly, the *P*-values obtained for marker genes are based on the assumption that the obtained cell clusters represent the biological ground truth. If one considers that there is uncertainty in the cluster assignment, the relationship between cluster assignment and marker gene detection must be accounted for in the statistical test. This relationship arises as both clusters and marker genes are typically determined based on the same gene expression data. The null hypothesis implicit in DE tests is that genes have the same distribution of expression values between the two groups. Yet, as the two groups are defined by the output of a clustering method in marker gene detection, there are differences in their gene expression profiles by design. We thus find significant marker genes even when clustering random data generated by splatter (Zappia *et al*, 2017) (see Appendix Supplementary Text S3). To obtain an appropriate measure of significance in clustered data, one can use a permutation test to account for the clustering step. This test is elaborated on in Appendix Supplementary Text S3. A recent differential expression tool also specifically addresses this issue (preprint: Zhang *et al*, 2018). With the current set-up, the *P*-values are often inflated, which can lead to an overestimation of the number of marker genes. However, the ranking of genes based on *P*-values is unaffected. Assuming the clustering is biologically meaningful, the top-ranked marker genes will still be the best marker gene candidates. In the first instance, we can loosely verify marker genes via visual inspection. We emphasize that inflated *P*-values occur specifically when defining cell-identity clusters via an unsupervised clustering approach. When instead determining cell-identity clusters via the expression of individual genes, the *P*-values can be interpreted as expected for all other genes. This univariate approach to cluster annotation, although common, is however not recommended outside specific cases (e.g. insulin in beta cells or haemoglobin in erythrocytes). Secondly, marker genes differentiate a cluster from others in the dataset and are thus dependent not only on the cell cluster, but also on the dataset composition. If the dataset composition does not accurately represent background gene expression, detected marker genes will be biased towards what is missing. This aspect must be considered especially when computing marker genes for datasets with low cellular diversity.

Recently, automated cluster annotation has become available. By directly comparing the gene expression profiles of annotated reference clusters to individual cells, tools such as scmap (Kiselev *et al*, 2018b) or Garnett (preprint: Pliner *et al*, 2019) can transfer annotations between the reference and the dataset. Thus, these methods can perform annotation and cluster assignment simultaneously, without the need for a data-driven clustering step. As cell type and state compositions differ between experimental conditions (Segerstolpe *et al*, 2016; Tanay & Regev, 2017), clustering based on reference data should not replace the data-driven approach.

The iteration of clustering, cluster annotation, re- or subclustering and re-annotation can be time-consuming. Automated cluster annotation methods offer a vast speedup of this process. However, there are benefits and limitations to automated and manual approaches that make it difficult to recommend one over the other. Increases in speed go together with compromises in flexibility. As mentioned above, reference atlases will not contain exactly the same cell identities as the dataset under investigation. Thus, one should not forgo marker gene calculation for manual annotation. Especially for large datasets that contain many clusters, the current best practice is a combination of both approaches. In the interest of speed, automated cell-identity annotation can be used to coarsely label cells and identify where subclustering may be needed. Subsequently, marker genes should be calculated for the dataset clusters and compared to known marker gene sets from the reference dataset or literature. For smaller datasets and datasets that lack reference atlases, manual annotation will suffice.

Pitfalls & recommendations:

- Do not use marker gene *P*-values to validate a cell-identity cluster, especially when the detected marker genes do not help to annotate the community. *P*-values may be inflated.
- Note that marker genes for the same cell-identity cluster may differ between datasets purely due to dataset cell type and state compositions.
- If relevant reference atlases exist, we recommend using automated cluster annotation combined with data-derived marker-gene-based manual annotation to annotate clusters.

Compositional analysis

At the cell level, we can analyse clustered data in terms of its compositional structure. Compositional data analysis revolves around the proportions of cells that fall into each cell-identity cluster. These proportions can change in response to disease. For example,

 © 2019 The Authors

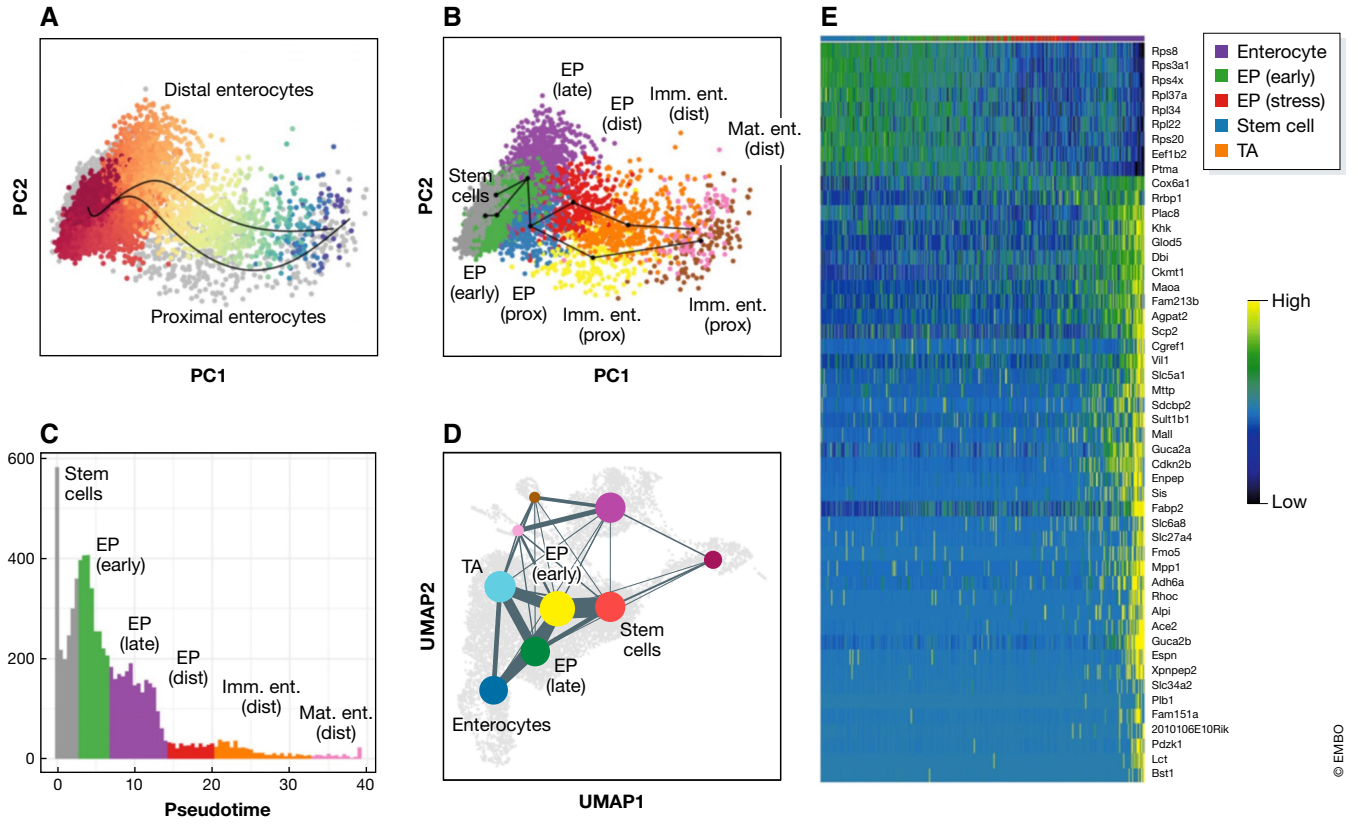


Figure 7. Trajectory analysis and graph abstraction of mouse intestinal epithelium data from Haber *et al* (2017).

(A) Distal and proximal enterocyte differentiation trajectories inferred by Slingshot. The Distal lineage is shown coloured by pseudotime from red to blue. Other cells in the dataset are grey. (B) Slingshot trajectories over clusters in PCA space. Clusters are abbreviated as follows: EP—enterocyte progenitors; Imm. Ent.—immature enterocytes; Mat. Ent.—mature enterocytes; Prox.—proximal; Dist.—distal. (C) Density over pseudotime for the distal enterocyte trajectory from Fig 7A. Colours represent the dominant cluster labels in each pseudotime bin. (D) Abstracted graph representation of the dataset projected onto a UMAP representation. Clusters are shown as coloured nodes. Clusters that appear in other trajectories are labelled for comparison. "TA" denotes transit amplifying cells. (E) Gene expression dynamics over pseudotime in a general enterocyte trajectory using the "GAM" R library.

salmonella infection has been shown to increase the proportion of enterocytes in the mouse intestinal epithelium (Haber *et al*, 2017).

Investigating compositional changes in single-cell data requires sufficient cell numbers to robustly assess cell-identity cluster proportions, and sufficient sample numbers to evaluate expected background variation in cell-identity cluster compositions. As appropriate datasets have only recently become available, dedicated tools are yet to be developed. In the mouse study mentioned above, cell-identity counts were modelled using a Poisson process, including the condition as a covariate, and the total number of cells detected as an offset. Here, a statistical test can be performed over the regression coefficient to assess whether a particular cell identity has significantly changed in frequency. However, tests over other cell identities in the same dataset are not independent of each other. If the proportion of one cell-identity cluster changes, the proportions of all others must have changed as well. Thus, one cannot assess whether the overall composition has significantly changed using this model. In the absence of dedicated tools, visual comparison of compositional data can be informative of changes in compositions between samples (Fig 6C). Future developments in this field will likely borrow from the mass cytometry (e.g. Tibshirani *et al*, 2002; Arvaniti & Claassen, 2017; Lun *et al*, 2017; Weber *et al*, 2018) or the microbiome literature (Gloor *et al*, 2017), where compositional data analysis has received more attention.

Pitfalls & recommendations:

- Consider that statistical tests over changes in the proportion of a cell-identity cluster between samples are dependent on one another.

Trajectory analysis

Trajectory inference

Cellular diversity cannot sufficiently be described by a discrete classification system such as clustering. The biological processes that drive the development of the observed heterogeneity are continuous processes (Tanay & Regev, 2017). Thus, in order to capture transitions between cell identities, branching differentiation processes, or gradual, unsynchronized changes in biological function, we require dynamic models of gene expression. This class of methods is known as trajectory inference (TI).

Trajectory inference methods interpret single-cell data as a snapshot of a continuous process. This process is reconstructed by finding paths through cellular space that minimize transcriptional changes between neighbouring cells (Fig 7A and B). The ordering of cells along these

© 2019 The Authors

paths is described by a *pseudotime* variable. While this variable is related to transcriptional distances from a root cell, it is often interpreted as a proxy for developmental time (Moignard *et al*, 2015; Haghverdi *et al*, 2016; Fischer *et al*, 2018; Griffiths *et al*, 2018).

Since Monocle (Trapnell *et al*, 2014) and Wanderlust (Bendall *et al*, 2014) established the TI field, the number of available methods has exploded. Currently available TI methods differ in the complexity of the paths that are modelled. Models range from simple linear or bifurcating trajectories, to complex graphs, trees, or multifurcating trajectories. In a recent comprehensive comparison of TI methods (Saelens *et al*, 2018), it was concluded that no individual method performs optimally for all types of trajectories. Instead, TI methods should be selected based on the complexity of the expected trajectory. The comparison revealed that Slingshot (Street *et al*, 2018) outperformed other methods for simple trajectories that range from linear to bi- and multifurcating models. If more complex trajectories are expected, PAGA (Wolf *et al*, 2019) was recommended by the authors. If the exact trajectory model is known, one can alternatively use more specialized methods to improve performance (Saelens *et al*, 2018). Generally, any inferred trajectory should be confirmed with an alternative method to avoid method bias.

In a typical workflow, TI methods are applied to reduced data or to corrected data when there is an inbuilt dimensionality reduction step. As multiple biological processes are typically occurring simultaneously within cells, it may be useful to regress out the biological effects of other processes to isolate the expected trajectory. For example, T cells may be undergoing cell cycle transitions during maturation (Buettner *et al*, 2015). Furthermore, as several top-performing TI methods rely on clustered data, TI is typically performed after clustering. Clusters in inferred trajectories may represent stable or metastable states (see "Metastable states"; Fig 7B and C). Subsequently, RNA velocities can be overlayed onto the trajectory to add directionality (La Manno *et al*, 2018).

Inferred trajectories do not have to represent biological processes. In the first instance, these only denote transcriptional similarity. Few TI methods include an evaluation of uncertainty in their model (Griffiths *et al*, 2018). Thus, further information is needed to validate whether a biological process was indeed captured. This information can come in the form of perturbation experiments, inferred regulatory gene dynamics, and support from RNA velocity.

Pitfalls & recommendations:

- We recommend using the Saelens *et al* (2018) review as a guide.
- Inferred trajectories do not have to represent a biological process. Further sources of evidence should be collected to interpret a trajectory.

Gene expression dynamics

One approach to garner support that an inferred trajectory is not the result of fitting transcriptional noise is to analyse the trajectory on the gene level. Genes that vary smoothly across pseudotime characterize the trajectory and can be used to identify the underlying biological process. Furthermore, this group of trajectory-associated genes is expected to contain genes that regulate the modelled process. Regulator genes help us understand how and why biological processes are triggered and represent potential drug targets (Gashaw *et al*, 2012).

While early approaches to find trajectory-associated genes involved DE testing between cell clusters along a trajectory (Haghverdi *et al*, 2016; Alpert *et al*, 2018), we now detect genes that vary across a trajectory by regressing gene expression against pseudotime. In order to enforce smooth variation of expression along this covariate, pseudotime is smoothed by fitting a spline or via an additional local regression step (e.g. loess). Regression frameworks differ in their noise model assumptions and the class of function used to describe the expression as a function of pseudotime. Potential regulatory genes are obtained by performing model selection for the genes' dependence on pseudotime. This DE test over pseudotime is confounded by the trajectory inference method in the same way that DE testing between clusters is confounded by the clustering method (see "Cluster annotation" section). Thus, *P*-values obtained in this set-up should not be regarded as an evaluation of significance.

Currently few dedicated gene temporal dynamics tools exist. BEAM is a tool integrated into the Monocle TI pipeline (Qiu *et al*, 2017a), which allows for detection of branch-specific gene dynamics. Outside of this pipeline, users can opt for LineagePulse (https://github.com/YosefLab/LineagePulse), which considers dropout noise but is still in development, or write their own testing framework using the limma package (Ritchie *et al*, 2015) or standard R libraries. An example of this can be found in the online Slingshot tutorial (Street *et al*, 2018) and in Fig 7E.

Given the few available tools, a best practice for investigating gene temporal dynamics cannot yet be determined. Exploratory investigation of gene dynamics is surely possible using all above methods. In future, Gaussian processes may provide a natural model to investigate gene temporal dynamics. Furthermore, testing for regulatory modules rather than individual genes would likely improve the signal-to-noise ratio and facilitate the biological interpretation.

Metastable states

Cell-level analysis of trajectories investigates cellular densities across pseudotime. Assuming that cells were sampled in an unbiased manner, dense regions along a trajectory indicate preferred transcriptomic states. When interpreting the trajectory as a temporal process, these dense regions may represent metastable states in, for example, development (Haghverdi *et al*, 2016). We can find these metastable states by plotting histograms of the pseudotime coordinate (Fig 7C).

Cell-level analysis unification

Clustering and trajectory inference represent two distinct views of single-cell data. These two views can be reconciled in coarse-grained graph representations. By representing single-cell clusters as nodes, and trajectories between the clusters as edges, one can represent both the static and dynamic nature of the data. This unification was proposed by the partition-based graph abstraction tool (PAGA; Fig 7D; Wolf *et al*, 2019). Using a statistical model for cell cluster interactions, PAGA places an edge between cluster nodes whose cells are more similar than expected. PAGA has been favourably compared to other TI methods in a recent review (Saelens *et al*, 2018). It was the only reviewed method able to cope with disconnected topologies and complex graphs containing cycles. This feature makes PAGA a helpful tool to visualize the topology of the entire dataset also for exploratory analysis.

 © 2019 The Authors

Gene-level analysis

While we have so far focused on gene-level analysis methods that characterize cellular structures, gene-level analysis of single-cell data has a broader scope. Differential expression testing, gene set analyses and gene regulatory network inference directly investigate molecular signals in the data. Rather than describing the cellular heterogeneity, these approaches use this heterogeneity as context in which gene expression is to be understood.

Differential expression testing

A common question asked of expression data is whether any genes are differentially expressed between two experimental conditions. DE testing is a well-documented problem that originates from bulk gene expression analysis (Scholtens & von Heydebreck, 2005). An advantage over bulk differential testing is that we can account for cellular heterogeneity in the single-cell setting by performing tests within cell-identity clusters. This set-up tells us how individual cell identities react transcriptionally under particular experimental conditions (Kang *et al*, 2018).

Although designed to answer the same question, bulk and single-cell DE tools differ methodologically. While bulk methods were developed to accurately estimate gene variance from few samples, single-cell data do not present this problem. On the other hand, single-cell data contain unique technical noise artefacts such as dropout, and high cell-to-cell variability (Hicks *et al*, 2017; Vallejos *et al*, 2017). These artefacts are taken into account in methods designed specifically for single-cell data (Kharchenko *et al*, 2014; Finak *et al*, 2015). Yet, a recent, large-scale comparison study of DE analysis has suggested that bulk DE testing packages perform comparably to the best-performing single-cell tools (Soneson & Robinson, 2018). Furthermore, when bulk tools are adapted to model single-cell data via introducing gene weights into the tests, these tools have been suggested to outperform their single-cell counterparts (Van den Berge *et al*, 2018). According to this comparison, the top-performing DE analysis tools are DESeq2 (Love *et al*, 2014) and EdgeR (Robinson *et al*, 2010) in combination with weights estimated by ZINB-wave (Risso *et al*, 2018). Independent comparison studies that include weighted bulk DE testing methods are required to confirm these results.

The improved performance of weighted bulk DE testing comes at the cost of computational efficiency. Given the trend of increasing cell numbers in single-cell experiments, algorithm runtime is becoming an increasingly important consideration in method choice. Thus, the single-cell tool MAST (Finak *et al*, 2015) represents a potent alternative to weighted bulk DE tools. MAST uses a hurdle model to account for dropout while modelling changes in gene expression dependent upon condition and technical covariates. It was the best-performing single-cell DE testing method in the aforementioned study (Soneson & Robinson, 2018), and outperformed bulk and single-cell methods in a small-scale comparison on a single dataset (Vieth *et al*, 2017). While MAST has a 10-fold to 100-fold faster runtime than weighted bulk methods (Van den Berge *et al*, 2018), a further 10-fold speedup can be achieved using limma–voom (Law *et al*, 2014). Although limma is a bulk DE testing method, limma–voom was shown to achieve comparable performance to MAST.

As uncorrected, measured data should be used for DE testing, accounting for confounding factors is crucial to robust estimation of differentially expressed genes. While DE testing tools typically allow the user the flexibility to incorporate confounders, users must be vigilant which variables are added to the model. For example, in most single-cell experimental set-ups the sample and condition covariates are confounded, since it is rarely possible to obtain a single sample under multiple conditions. If we incorporate both the sample and condition covariates into the model, the variability associated with these covariates can no longer unambiguously be assigned. Thus, when testing over condition, we cannot include the sample covariate in the model in the given form. When correcting for multiple categorical batch covariates, it becomes increasingly difficult to visually spot confounding groups of covariates. In this situation, it is helpful to test whether the model design matrix is full rank. Even when design matrices are not full rank, DE testing tools will often adapt the matrix and run without outputting a warning. This will not deliver the intended results.

In the scenario we describe here, the condition covariate is determined in the experimental set-up. Thus, a DE test over this covariate (within the same cluster) is independent of the clustering procedure. This set-up distinguishes DE testing over conditions and DE testing over clusters. Obtained *P*-values for DE tests over conditions represent the expected measures of significance and must be corrected for multiple testing. To reduce the multiple testing burden, transcripts that may not be of interest can be excluded from the dataset. While pseudogenes or non-coding RNAs can be informative (An *et al*, 2017), they are often ignored in the analysis.

Pitfalls & recommendations:

- DE testing should not be performed on corrected data (denoised, batch corrected, etc.), but instead on measured data with technical covariates included in the model.
- Users should not rely on DE testing tools to correct models with confounded covariates. Model specification should be performed carefully ensuring a full-rank design matrix.
- We recommend using MAST or limma for DE testing.

Gene set analysis

Gene-level analysis methods often produce long lists of candidate genes that are difficult to interpret. For example, thousands of genes may be differentially expressed between treated and control cells. We can facilitate the interpretation of these results by grouping the genes into sets based on shared characteristics and testing whether these characteristics are overrepresented in the candidate gene list.

Gene set information can be found in curated label databases for various applications. To interpret DE results, we typically group genes based on involvement in common biological processes. Biological process labels are stored in databases such as MSigDB (Liberzon *et al*, 2011), the Gene Ontology (Ashburner *et al*, 2000; The Gene Ontology Consortium, 2017), or the pathway databases KEGG (Kanehisa *et al*, 2017) and Reactome (Fabregat *et al*, 2018). Enrichment of annotations on the gene list can be tested using a vast array of tools, which are reviewed and compared in Huang *et al* (2009) and Tarca *et al* (2013).

A recent development in the single-cell analysis field is the use of paired gene labels to perform ligand–receptor analysis. Here, interaction between cell clusters is inferred from the expression of receptors and their cognate ligands. Ligand–receptor pair labels can be

© 2019 The Authors

obtained from the recent CellPhoneDB (Vento-Tormo *et al*, 2018) and used to interpret the highly expressed genes across clusters using statistical models (Zepp *et al*, 2017; Zhou *et al*, 2017; Cohen *et al*, 2018; Vento-Tormo *et al*, 2018).

Gene regulatory networks

Genes do not function independently. Instead, the expression level of a gene is determined by a complex interplay of regulatory interactions with other genes and small molecules. Uncovering these regulatory interactions is the goal of gene regulatory network (GRN) inference methods.

Gene regulatory network inference is performed based on measurements of gene co-expression such as correlation, mutual information, or via regression models (Chen & Mar, 2018). If two genes show a co-expression signal even when all other genes are taken into account as potential confounders, these genes are said to have a causal regulatory relationship. Inferring gene regulatory relationships is related to the detection of trajectory-associated regulatory genes. Indeed, several single-cell GRN inference methods use trajectories with mechanistic differential equation models (Ocone *et al*, 2015; Matsumoto *et al*, 2017).

While there exist GRN inference methods that were specifically developed for scRNA-seq data (SCONE: Matsumoto *et al*, 2017; PIDC: Chan *et al*, 2017; SCENIC: Aibar *et al*, 2017), a recent comparison has shown both bulk and single-cell methods to perform poorly on these data (Chen & Mar, 2018). GRN inference methods may still offer valuable insights to identify causal regulators of biological processes, yet we recommend that these methods be used with care.

Pitfalls & recommendations:

- Users should be wary of uncertainty in the inferred regulatory relationships. Modules of genes that are enriched for regulatory relationships will be more reliable than individual edges.

Analysis platforms

Single-cell analysis workflows are collations of independently developed tools. To facilitate the movement of data between these tools, single-cell platforms have been developed around consistent data formats. These platforms provide a basis for the construction of analysis pipelines. Currently available platforms exist on the command line in R (McCarthy *et al*, 2017; Butler *et al*, 2018) or Python (Wolf *et al*, 2018), and as local applications (Patel, 2018; preprint: Scholz *et al*, 2018) or Web servers (Gardeux *et al*, 2017; Zhu *et al*, 2017) with graphical user interfaces (GUIs). An overview of platforms is available in Zhu *et al* (2017) and Zappia *et al* (2018).

Among command line platforms, Scater (McCarthy *et al*, 2017) and Seurat (Butler *et al*, 2018) easily interface with the large variety of analysis tools available via the R Bioconductor project (Huber *et al*, 2015). Scater has a particular strength in QC and pre-processing, while Seurat is arguably the most popular and comprehensive platform, which includes a large array of tools and tutorials. A recent addition to this group is scanpy (Wolf *et al*, 2018), a growing Python-based platform, which exhibits improved scaling to larger numbers of cells. It leverages the increasing number of tools written in Python, which is particularly popular for machine learning applications.

Graphical user interface platforms enable non-expert users to build single-cell analysis workflows. Users are often guided through prescribed workflows that facilitate the analysis, but also limit user flexibility. These platforms are especially useful for exploratory analysis. Platforms such as Granatum (Zhu *et al*, 2017) and ASAP (Gardeux *et al*, 2017) differ in the tools they integrate, with Granatum including the larger variety of methods. As Web servers, these two platforms are readily available, yet computational infrastructure will limit their ability to scale to large datasets. For example, ASAP was tested on a dataset of only 92 cells. Alternatives to the Web-based GUI platforms are packages such as FASTGenomics (preprint: Scholz *et al*, 2018), iSEE (Rue-Albrecht *et al*, 2018), IS-CellR (Patel, 2018), and Granatum run on a local server. These are platforms and GUI wrappers that can scale with the locally available computational power. In future, the ongoing development of the Human Cell Atlas portals (https://www.humancellatlas.org/data-sharing) will lead to more powerful visual data exploration tools that scale to large cell numbers.

Conclusions and outlook

We have reviewed the steps of a typical scRNA-seq analysis workflow and implemented these in a case study tutorial (https://www.github.com/theislab/single-cell-tutorial). The tutorial was designed to follow current best practices as determined by available method comparisons. While aggregating individual best-practice tools does not guarantee an optimal pipeline, we hope that our workflow represents a current snapshot of the state of the art in the single-cell analysis field. It thus provides a suitable entry point into this field for newcomers and contributes to the efforts of the Human Cell Atlas to establish best practices in scRNA-seq analysis (preprint: Regev *et al*, 2018). It should be noted that available method comparisons necessarily lag behind the latest method developments. Thus, we have mentioned new developments that have not yet been independently evaluated where possible. With the future development of new and better tools, and further comparative studies, the individual tool recommendations presented here will require updates, yet the general considerations regarding the stages of data processing should remain the same.

Two avenues of development that are of particular interest due to their potential for disruption to analysis pipelines are deep learning workflows and single-cell omic integration. Due to its flexibility to scale to large data, deep learning has revolutionized fields from computer vision to natural language processing, and is starting to have a strong impact in genomics (Webb, 2018). First applications to scRNA-seq are starting to emerge from dimensionality reduction to denoising (e.g. scVis: Ding *et al*, 2018; scGen: preprint: Lotfollahi *et al*, 2018; DCA: Eraslan *et al*, 2019). Recently, deep learning has been used to produce an embedded workflow that can fit the data, denoise it and perform downstream analysis such as clustering and differential expression within the framework of the model (scVI: Lopez *et al*, 2018). In this set-up, it is possible to include noise and batch effect estimates into downstream statistical tests while preserving accurate estimates of variation in the data. Integrated

 © 2019 The Authors

modelling approaches such as this have the potential to replace current pipelines, which are often an agglomeration of individual tools.

As single-cell omic technologies improve, the need for integrated omic analysis pipelines will grow (Tanay & Regev, 2017). Future single-cell platforms will have to be able to deal with different data sources such as DNA methylation (Smallwood *et al*, 2014), chromatin accessibility (Buenrostro *et al*, 2015), or protein abundance (Stoeckius *et al*, 2017), and include tools that integrate these modalities. For this set-up, it will no longer be possible to use only a single read or count matrix, which we use as the starting point of our tutorial. However, platforms are already adapting to multi-modal data structures for the integration of RNA velocity, which is calculated from unspliced and spliced read data (La Manno *et al*, 2018). Single-cell multi-omic integration can occur via consensus clustering approaches, multi-omic factor analysis (Argelaguet *et al*, 2018), or multi-omic gene regulatory network inference (Colomé-Tatché & Theis, 2018). Analysis workflows with these capabilities will be the next stage of development. We envisage that such multi-omic analysis workflows will build upon the foundation we have laid for scRNA-seq.

Expanded View for this article is available online.

Acknowledgements
Special thanks to Leander Dony, who debugged, updated, and tested the case study to work with the latest methods. Furthermore, we would like to thank the many people who proofread the case study notebook and the manuscript and improved it with their comments and expertise. For this, we acknowledge the input of Maren Buttner, David Fischer, Alex Wolf, Lukas Simon, Luis Ospina-Forero, Sophie Tritschler, Niklas Koehler, Goekcen Eraslan, Benjamin Schubert, Meromit Singer, Dana Pe'er, and Rahul Satija. Special thanks for this also to the anonymous reviewers of the manuscript and the editor, Thomas Lemberger, for their thorough, constructive, and extensive comments. The case study notebook was tested and improved by the early adopters Marius Lange, Hananeh Aliee, Subarna Palit and Lisa Thiergart. Volker Bergen and Alex Wolf also contributed to the workflow by making scanpy adaptations. The choice of dataset to optimally show all aspects of the analysis workflow was facilitated by the kind input from Adam Haber and Aviv Regev. This work was supported by the BMBF grant# 01IS18036A and grant# 01IS18053A, by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17, by the Helmholtz Association (Incubator grant sparse2big, grant # ZT-I-0007) and by the Chan Zuckerberg Initiative DAF (advised fund of Silicon Valley Community Foundation, 182835).

Author contributions
MDL reviewed the literature and wrote the paper, and FJT supervised the work and critically reviewed the manuscript. Both authors read and approved the final paper.

Conflict of interest
The authors declare that they have no conflict of interest.

References

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J *et al* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 14: 1083 – 1086

Alpert A, Moore LS, Dubovik T, Shen-Orr SS (2018) Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat Methods* 15: 267 – 270

An Y, Furber KL, Ji S (2017) Pseudogenes regulate parental gene expression via ceRNA network. *J Cell Mol Med* 21: 185 – 192

Andrews TS, Hemberg M (2018) False signals induced by single-cell imputation. *F1000Res* 7: 1740

Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, Tsitsiridis G, Graf E, Strom TM, Eickelberg O *et al* (2019) An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun* 10: 963

Angerer P, Simon L, Tritschler S, Alexander Wolf F, Fischer D, Theis FJ (2017) Single cells make big data: new challenges and opportunities in transcriptomics. *Curr Opin Syst Biol* 4: 85 – 91

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 14: e8124

Arvaniti E, Claassen M (2017) Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun* 8: 14825

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25 – 29

Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M *et al* (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* 174: 1293 – 1308.e36

Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 37: 38 – 44

Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157: 714 – 725

Blasi T, Buettner F, Strasser MK, Marr C, Theis FJ (2017) cgCorrect: a method to correct for confounding cell-cell variation due to cell growth in single-cell transcriptomics. *Phys Biol* 14: 036001

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008: P10008

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC *et al* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 10: 1093 – 1095

Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM (2018) The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360 : eaar5780

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523: 486 – 490

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33: 155 – 160

Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 18: 212

© 2019 The Authors

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36: 411 – 420

Buttner M, Miao Z, Wolf A, Teichmann SA, Theis FJ (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 16: 43 – 49

Chan TE, Stumpf MPH, Babtie AC (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst* 5: 251 – 267.e3

Chen S, Mar JC (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* 19: 232

Chung NC, Storey JD (2015) Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* 31: 545 – 554

Clevers H, Rafelski S, Elowitz M, Klein AM, Shendure J, Trapnell C, Lein E, Lundberg E, Uhlén M, Martinez-Arias A *et al* (2017) What is your conceptual definition of 'cell type' in the context of a mature organism? *Cell Syst* 4: 255 – 259

Cohen M, Giladi A, Gorki A-D, Solodkin DG, Zada M, Hladik A, Miklosi A, Salame T-M, Halpern KB, David E *et al* (2018) Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* 175: 1031 – 1044

Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci USA* 102: 7426 – 7431

Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, Dudoit S, Yosef N (2019) Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst* 8: 315 – 328

Colomé-Tatché M, Theis FJ (2018) Statistical single cell multi-omics integration. *Curr Opin Syst Biol* 7: 54 – 59

DePasquale EAK, Schnell DJ, Valiente I, Blaxall BC, Grimes HL, Singh H, Salomonis N (2018) DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq Doublets. *bioRxiv* https://doi.org/10.1101/364810 [PREPRINT]

van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D *et al* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell* 174: 716 – 729.e27

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14: 671 – 683

Ding J, Condon A, Shah SP (2018) Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun* 9: 2002

Duò A, Robinson MD, Soneson C (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 7: 1141 .

Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ (2019) Single cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 10: 390

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B *et al* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res* 46: D649 – D655

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, Juliana McElrath M, Prlic M *et al* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16: 278

Fischer DS, Theis FJ, Yosef N (2018) Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Res* 46: e119

Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M (2018) Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Res* 7: 1297

Gardeux V, David FPA, Shajkofci A, Schwalie PC, Deplancke B (2017) ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* 33: 3123 – 3125

Gashaw I, Ellinghaus P, Sommer A, Asadullah K (2012) What makes a good drug target?. *Drug Discov Today* 17: S24 – S30

Gehring J, Park JH, Chen S, Thomson M, Pachter L (2018) Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. *bioRxiv* https://doi.org/10.1101/315333 [PREPRINT]

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: and this is not optional. *Front Microbiol* 8: 2224

Griffiths JA, Scialdone A, Marioni JC (2018) Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol Syst Biol* 14: e8046

Grün D, Kester L, van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. *Nat Methods* 11: 637

Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y *et al* (2017) A single-cell survey of the small intestinal epithelium. *Nature* 551: 333 – 339

Haghverdi L, Buettner F, Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31: 2989 – 2998

Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* 13: 845 – 848

Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36: 421 – 427

Heimberg G, Bhatnagar R, El-Samad H, Thomson M (2016) Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst* 2: 239 – 250

Hicks SC, Townes FW, Teng M, Irizarry RA (2017) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19: 562 – 578

Hie BL, Bryson B, Berger B (2018) Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv* https://doi.org/10.1101/371179 [PREPRINT]

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1 – 13

Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 15: 539 – 542

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T *et al* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 12: 115 – 121

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* 17: 29

Johnson WE, Li C, Rabinovic A (2006) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118 – 127

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45: D353 – D361

Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM *et al* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 36: 89 – 94

 © 2019 The Authors

Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods* 11: 740 – 742

Kim T, Chen IR, Lin Y, Wang AY-Y, Yang JYH, Yang P (2018) Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform* https://doi.org/10.1093/bib/bby076

Kiselev VY, Yiu A, Hemberg M (2018b) Scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 15: 359 – 362

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161: 1187 – 1201

Korsunsky I, Fan J, Slowikowski K, Zhang F, Wei K (2018) Fast, sensitive, and accurate integration of single cell data with Harmony. *bioRxiv* https://doi.org/10.1101/461954 [PREPRINT]

Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, Schneider RK, Wagers AJ, Ebert BL, Regev A (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* 25: 1860 – 1872

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A *et al* (2018) RNA velocity of single cells. *Nature* 560: 494 – 498

Law CW, Chen Y, Shi W, Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15: R29

Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER *et al* (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162: 184 – 197

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2009) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26: 493 – 500

Li WV, Li JJ (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 9: 997

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739 – 1740

Liu Y, Wang T, Zheng D (2018) RISC: robust integration of single-cell RNA-seq datasets with different extents of cell cluster overlap. *bioRxiv* https://doi.org/10.1101/483297 [PREPRINT]

Lopez R, Regier J, Cole MB, Jordan M, Yosef N (2018) Deep generative modeling fro single-cell transcriptomics. *Nat Methods* 15: 1053 – 1058

Lotfollahi M, Wolf FA, Theis FJ (2018) Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. *bioRxiv* https://doi.org/10.1101/478503 [PREPRINT]

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Lun ATL, Bach K, Marioni JC (2016a) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 17: 75

Lun ATL, McCarthy DJ, Marioni JC (2016b) A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Res* 5: 2122

Lun ATL, Richard AC, Marioni JC (2017) Testing for differential abundance in mass cytometry data. *Nat Methods* 14: 707 – 709

Lun ATL (2018) Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *bioRxiv* https://doi.org/10.1101/404962 [PREPRINT]

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9: 2579 – 2605

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM *et al* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161: 1202 – 1214

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp 281 – 297. Oakland, CA.

Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 33: 2314 – 2321

Mayer C, Hafemeister C, Bandler RC, Machold R, Batista Brito R, Jaglin X, Allaway K, Butler A, Fishell G, Satija R (2018) Developmental diversification of cortical inhibitory interneurons. *Nature* 555: 457 – 462

McCarthy DJ, Campbell KR, Lun ATL, Wills QF (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33: 1179 – 1186

McDavid A, Finak G, Gottardo R (2016) The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat Biotechnol* 34: 591 – 593

McGinnis CS, Murrow LM, Gartner ZJ (2018) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* 8: 329 – 337

McInnes L, Healy J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv* https://arxiv.org/abs/1802.03426v2 [PREPRINT]

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E *et al* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* 33: 269 – 276

Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM, Villoria J *et al* (2018) A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 560: 319 – 324

Moon KR, Stanley JS, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S (2018) Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr Opin Syst Biol* 7: 36 – 46

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 69: 026113

Ocone A, Haghverdi L, Mueller NS, Theis FJ (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* 31: i89 – i96

Pachter L (2011) Models for transcript quantification from RNA-Seq. *arXiv* https://arxiv.org/abs/1104.3889v2 [PREPRINT]

Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I (2018) zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* 7: giy059

Park JE, Polanski K, Meyer K, Teichmann SA (2018) Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape. *bioRxiv* https://doi.org/10.1101/397042 [PREPRINT]

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL *et al* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344: 1396 – 1401

Patel MV (2018) iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. *Bioinformatics* 34: 4305 – 4306

Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos Mag* 2: 559 – 572

© 2019 The Authors

Plass M, Solana J, Alexander Wolf F, Ayoub S, Misios A, Glažar P, Obermayer B, Theis FJ, Kocks C, Rajewsky N (2018) Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360: eaaq1723

Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB (2018) A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 560: 377 – 381

Pliner HA, Shendure J, Trapnell C (2019) Supervised classification enables rapid annotation of cell atlases. *bioRxiv* https://doi.org/10.1101/538652 [PREPRINT]

Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C (2017a) Single-cell mRNA quantification and differential analysis with census. *Nat Methods* 14: 309 – 315

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M *et al* (2017) Science forum: the human cell atlas. *Elife* 6: e27041

Regev A, Teichmann S, Rozenblatt-Rosen O, Stubbington M, Ardlie K, Amit I, Arlotta P, Bader G, Benoist C, Biton M *et al* (2018) The human cell atlas white paper. *arXiv* https://arxiv.org/abs/1810.05192 [PREPRINT]

Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys* 74: 016110

Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 9: 284

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139 – 140

Rostom R, Svensson V, Teichmann SA, Kar G (2017) Computational approaches for interpreting scRNA-seq data. *FEBS Lett* 591: 2213 – 2225

Rue-Albrecht K, Marini F, Soneson C, Lun ATL (2018) iSEE: interactive summarizedexperiment explorer. *F1000Res* 7: 741

Saelens W, Cannoodt R, Todorov H, Saeys Y (2018) A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 37: 547 – 554

Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S *et al* (2018) Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 174: 1015 – 1030.e16

Scholtens D, von Heydebreck A (2005) Analysis of Differential Gene Expression Studies. In *Statistics for Biology and Health*, pp 229 – 248.

Scholz CJ, Biernat P, Becker M, Baßler K, Günther P, Balfer J, Dickten H, Flöer L, Heikamp K, Angerer P *et al* (2018) FASTGenomics: An analytical ecosystem for single-cell RNA sequencing data. *bioRxiv* https://doi.org/10.1101/272476 [PREPRINT]

Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK *et al* (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 24: 593 – 607

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11: 817 – 820

Soneson C, Robinson MD (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* 15: 255 – 261

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14: 865 – 868

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom* 19: 477

Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14: 381 – 387

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group & Principal investigators (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562: 367 – 372

Tanay A, Regev A (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541: 331 – 338

Tarca AL, Bhatti G, Romero R (2013) A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* 8: e79217

The Gene Ontology Consortium (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res* 45: D331 – D338

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 99: 6567 – 6572

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32: 381 – 386

Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 14: 565 – 571

Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, Robinson MD, Dudoit S, Clement L (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol* 19: 24

Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park J, Stephenson E, Polański K, Goncalves A *et al* (2018) Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563: 347 – 353

Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I (2017) powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* 33: 3486 – 3488

Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I (2019) A systematic evaluation of single cell RNA-seq analysis pipelines: library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies. *bioRxiv* https://doi.org/10.1101/583013 [PREPRINT]

Wagner A, Regev A, Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* 34: 1145 – 1160

Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360: 981 – 987

Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 14: 414 – 416

Wattenberg M, Viégas F, Johnson I (2016) How to use t-SNE effectively. *Distill*. https://doi.org/10.23915/distill.00002

Webb S (2018) Deep learning for biology. *Nature* 554: 555 – 557

Weber LM, Robinson MD (2016) Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* 89: 1084 – 1096

 © 2019 The Authors

Weber LM, Nowicka M, Soneson C & Robinson MD (2018) diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun Biol* 2: 183

Weinreb C, Wolock S, Klein AM (2018) SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* 34: 1246 – 1248

Welch J, Kozareva V, Ferreira A, Vanderburg C (2018) Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv* https://doi.org/10.1101/459891 [PREPRINT]

Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19: 15

Wolf FA, Hamey F, Plass M, Solana J, Dahlin JS, Gottgens B, Rajewsky N, Simon L, Theis FJ (2019) Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 20: 59

Wolock SL, Lopez R, Klein AM (2019) Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* 8: 281 – 291.

Yip SH, Sham PC, Wang J (2018) Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform*: bby011

Young MD, Behjati S (2018) SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv* https://doi.org/10.1101/303727 [PREPRINT]

Zappia L, Phipson B, Oshlack A (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 18: 174

Zappia L, Phipson B, Oshlack A (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* 14: e1006245

Zeisel A, Hochgerner H, Lonnerberg P, Johnsson A, Memic F, van der Zwan J, Haring M, Braun E, Borm L, La Manno G *et al* (2018) Molecular architecture of the mouse nervous system. *Cell* 174: 999 – 1014

Zepp JA, Zacharias WJ, Frank DB, Cavanaugh CA, Zhou S, Morley MP, Morrisey EE (2017) Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. *Cell* 170: 1134 – 1148.e10

Zhang JM, Kamath GM, Tse DN (2018) Towards a post-clustering test for differential expression. *bioRxiv* https://doi.org/10.1101/463265 [PREPRINT]

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J *et al* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8: 14049

Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, Huang S (2017) Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. *Sci Rep* 7: 8815

Zhu X, Wolfgruber TK, Tasato A, Arisdakessian C, Garmire DG, Garmire LX (2017) Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med* 9: 108

Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 65: 631 – 643.e4

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

A.8 Proteomics

Recommended reading for Chapter 8 (listed in Section 8.9).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Aebersold and Mann (2016) — mass-spectrometric exploration of proteome structure and function; the field in one review.
  DOI · via SU library (SU login)

  Mass-spectrometric exploration of proteome structure and function. Aebersold & Mann (2016). *Nature*. doi:10.1038/nature19949

  > Numerous biological processes are concurrently and coordinately active in every living cell. Each of them encompasses synthetic, catalytic and regulatory functions that are, almost always, carried out by proteins organized further into higher-order structures and networks. For decades, the structures and functions of selected proteins have been studied using biochemical and biophysical methods. However, the properties and behaviour of the proteome as an integrated system have largely remained elusive. Powerful mass-spectrometry-based technologies now provide unprecedented insights into the composition, structure, function and control of the proteome, shedding light on complex biological processes and phenotypes.

- Cox and Mann (2008) — MaxQuant; how a complete shotgun proteomics pipeline identifies and quantifies proteins.
  DOI · via SU library (SU login)

  MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Cox & Mann (2008). *Nature Biotechnology*. doi:10.1038/nbt.1511

  > Efficient analysis of very large amounts of raw data for peptide identification and protein quantification is a principal challenge in mass spectrometry (MS)-based proteomics. Here we describe MaxQuant, an integrated suite of algorithms specifically developed for high-resolution, quantitative MS data. Using correlation analysis and graph theory, MaxQuant detects peaks, isotope clusters and stable amino acid isotope-labeled (SILAC) peptide pairs as three-dimensional objects in m/z, elution time and signal intensity space. By integrating multiple mass measurements and correcting for linear and nonlinear mass offsets, we achieve mass accuracy in the p.p.b. range, a sixfold increase over standard techniques. We increase the proportion of identified fragmentation spectra to 73% for SILAC peptide pairs via unambiguous assignment of isotope and missed-cleavage state and individual mass precision. MaxQuant automatically quantifies several hundred thousand peptides per SILAC-proteome experiment and allows statistically robust identification and quantification of >4,000 proteins in mammalian cell lysates.

- Elias and Gygi (2007) — the target–decoy strategy; how false discoveries are controlled in proteomics.
  DOI · via SU library (SU login)

  Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Elias & Gygi (2007). *Nature Methods*. doi:10.1038/nmeth1019

  > Liquid chromatography and tandem mass spectrometry (LC-MS/MS) has become the preferred method for conducting large-scale surveys of proteomes. Automated interpretation of tandem mass spectrometry (MS/MS) spectra can be problematic, however, for a variety of reasons. As most sequence search engines return results even for 'unmatchable' spectra, proteome researchers must devise ways to distinguish correct from incorrect peptide identifications. The target-decoy search strategy represents a straightforward and effective way to manage this effort. Despite the apparent simplicity of this method, some controversy surrounds its successful application. Here we clarify our preferred methodology by addressing four issues based on observed decoy hit frequencies: (i) the major assumptions made with this database search strategy are reasonable; (ii) concatenated target-decoy database searches are preferable to separate target and decoy database searches; (iii) the theoretical error associated with target-decoy false positive (FP) rate measurements can be estimated; and (iv) alternate methods for constructing decoy databases are similarly effective once certain considerations are taken into account.

- Ludwig et al. (2018) — data-independent acquisition (SWATH-MS), a tutorial.
full text below · DOI · open access (CC BY 4.0)

  Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Ludwig et al. (2018). *Molecular Systems Biology*. doi:10.15252/msb.20178126

  > Many research questions in fields such as personalized medicine, drug screens or systems biology depend on obtaining consistent and quantitatively accurate proteomics data from many samples. SWATH-MS is a specific variant of data-independent acquisition (DIA) methods and is emerging as a technology that combines deep proteome coverage capabilities with quantitative consistency and accuracy. In a SWATH-MS measurement, all ionized peptides of a given sample that fall within a specified mass range are fragmented in a systematic and unbiased fashion using rather large precursor isolation windows. To analyse SWATH-MS data, a strategy based on peptide-centric scoring has been established, which typically requires prior knowledge about the chromatographic and mass spectrometric behaviour of peptides of interest in the form of spectral libraries and peptide query parameters. This tutorial provides guidelines on how to set up and plan a SWATH-MS experiment, how to perform the mass spectrometric measurement and how to analyse SWATH-MS data using peptide-centric scoring. Furthermore, concepts on how to improve SWATH-MS data acquisition, potential trade-offs of parameter settings and alternative data analysis strategies are discussed.

- Yilmaz et al. (2024) — Casanovo; de novo peptide sequencing as translation with a transformer.
full text below · DOI · open access (CC BY 4.0)

  Sequence-to-sequence translation from mass spectra to peptides with a transformer model. Yilmaz et al. (2024). *Nature Communications*. doi:10.1038/s41467-024-49731-x

  > A fundamental challenge in mass spectrometry-based proteomics is the identification of the peptide that generated each acquired tandem mass spectrum. Approaches that leverage known peptide sequence databases cannot detect unexpected peptides and can be impractical or impossible to apply in some settings. Thus, the ability to assign peptide sequences to tandem mass spectra without prior information-de novo peptide sequencing-is valuable for tasks including antibody sequencing, immunopeptidomics, and metaproteomics. Although many methods have been developed to address this problem, it remains an outstanding challenge in part due to the difficulty of modeling the irregular data structure of tandem mass spectra. Here, we describe Casanovo, a machine learning model that uses a transformer neural network architecture to translate the sequence of peaks in a tandem mass spectrum into the sequence of amino acids that comprise the generating peptide. We train a Casanovo model from 30 million labeled spectra and demonstrate that the model outperforms several state-of-the-art methods on a cross-species benchmark dataset. We also develop a version of Casanovo that is fine-tuned for non-enzymatic peptides. Finally, we demonstrate that Casanovo's superior performance improves the analysis of immunopeptidomics and metaproteomics experiments and allows us to delve deeper into the dark proteome.

- Uhlén et al. (2015) — the tissue-based map of the human proteome (Human Protein Atlas).
DOI · via SU library (SU login)

  Tissue-based map of the human proteome. Uhlén et al. (2015). *Science*. doi:10.1126/science.1260419

  > Resolving the molecular details of proteome variation in the different tissues and organs of the human body will greatly increase our knowledge of human biology and disease. Here, we present a map of the human tissue proteome based on an integrated omics approach that involves quantitative transcriptomics at the tissue and organ level, combined with tissue microarray-based immunohistochemistry, to achieve spatial localization of proteins down to the single-cell level. Our tissue-based analysis detected more than 90% of the putative protein-coding genes. We used this approach to explore the human secretome, the membrane proteome, the druggable proteome, the cancer proteome, and the metabolic functions in 32 different tissues and organs. All the data are integrated in an interactive Web-based database that allows exploration of individual proteins, as well as navigation of global expression patterns, in all major tissues and organs in the human body.

The open-access papers are reproduced below under their original licences.

Review

OPEN ACCESS

molecular systems biology

Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial

Christina Ludwig[1,*,†], Ludovic Gillet[2,†], George Rosenberger[2,3], Sabine Amon[2], Ben C Collins[2] & Ruedi Aebersold[2,4]

Abstract

Many research questions in fields such as personalized medicine, drug screens or systems biology depend on obtaining consistent and quantitatively accurate proteomics data from many samples. SWATH-MS is a specific variant of data-independent acquisition (DIA) methods and is emerging as a technology that combines deep proteome coverage capabilities with quantitative consistency and accuracy. In a SWATH-MS measurement, all ionized peptides of a given sample that fall within a specified mass range are fragmented in a systematic and unbiased fashion using rather large precursor isolation windows. To analyse SWATH-MS data, a strategy based on peptide-centric scoring has been established, which typically requires prior knowledge about the chromatographic and mass spectrometric behaviour of peptides of interest in the form of spectral libraries and peptide query parameters. This tutorial provides guidelines on how to set up and plan a SWATH-MS experiment, how to perform the mass spectrometric measurement and how to analyse SWATH-MS data using peptide-centric scoring. Furthermore, concepts on how to improve SWATH-MS data acquisition, potential trade-offs of parameter settings and alternative data analysis strategies are discussed.

Keywords data-independent acquisition; mass spectrometry; quantitative proteomics; SWATH-MS; systems biology

DOI 10.15252/msb.20178126 | Received 28 November 2017 | Revised 11 May 2018 | Accepted 15 May 2018

Mol Syst Biol. (2018) 14: e8126

Introduction

Over the last decades, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) has become the technology of choice for the high-throughput characterization of proteins and proteomes (Aebersold & Mann, 2016). Recent developments in the field have moved beyond enumerating the proteins, peptides or post-translational modifications detected in one or few samples towards delivering high quality and consistent quantification in large-scale projects that comprise 100s of samples. Especially in areas such as personalized medicine, biomarker research, drug screens, genetic association studies or systems biology, large numbers of individuals, conditions and perturbations need to be investigated to draw meaningful biological conclusions. For this purpose, the large data matrices generated must be as reproducible, complete and accurate as possible. In order to address these needs, several different proteomic strategies have been developed over the last years.

An emerging strategy, and the focus of this tutorial, is sequential window acquisition of all theoretical mass spectra (SWATH-MS), which was described by Gillet *et al* (2012). For a detailed introduction into the history, the basic principles as well as the general advantages and limitations of SWATH-MS (also summarized in Table 1), we refer to the Appendix. Briefly, for a SWATH-MS measurement, typically non-labelled protein samples are digested with trypsin and the resulting peptides are analysed by liquid chromatography coupled to a tandem mass spectrometer operating in the so-called data-independent acquisition (DIA) mode. In this mode, all ionized compounds of a given sample that fall within a specified mass range are fragmented in a systematic and unbiased fashion. Figure 1A–C shows the DIA scheme described as the initial implementation of SWATH-MS (Gillet *et al*, 2012), using 32 consecutive, slightly overlapping precursor isolation windows, with a width of 25 m/z each. Depending on sample complexity, this acquisition scheme will lead to the co-fragmentation of many co-eluting peptides concurrently selected in the precursor ion window and ultimately to highly multiplexed and complex fragment ion spectra (Fig 1D). To deal with this complexity, Gillet *et al* proposed a novel data analysis strategy based on peptide-centric scoring, which relies on querying chromatographic and mass spectrometric coordinates of the proteins and peptides of interest in form of so-called peptide query parameters (PQPs). PQPs are typically derived from previously generated spectral libraries.

In addition to the described SWATH-MS method, a wealth of other DIA schemes and alternative data analysis strategies has been

1 Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), Technical University of Munich (TUM), Freising, Germany
2 Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
3 Department of Systems Biology, Columbia University, New York, NY, USA
4 Faculty of Science, University of Zurich, Zurich, Switzerland
*Corresponding author. Tel: +49 8161 71 6130; E-mail: tina.ludwig@tum.de
†These authors contributed equally to this work

© 2018 The Authors. Published under the terms of the CC BY 4.0 license

Table 1. Advantages and limitations of SWATH-MS in comparison with data-dependent (DDA) and targeted (SRM, PRM) proteomics.

	Data independent acquisition-based SWATH-MS	Data-dependent acquisition (DDA)	Targeted acquisition (SRM and PRM)
Ease of data acquisition	** Easy, requires definition of mass range to cover, precursor isolation window width and number of MS2 scans per cycle	*** Easiest, default setup on most mass spectrometers, requires definition of TopN method, MS2 trigger threshold and dynamic exclusion time	* Hardest, requires generation and optimization of targeted peptide assays for data acquisition
Ease of data analysis	* Currently hardest, requires peptide query parameters, sophisticated software tools, large informatics resources	*** Currently easiest, multitude of pipelines available	** Currently easy, several software tools developed for manual and automated analysis
Breadth of protein and peptide detection/multiplexing	*** 10,000s of peptides per MS injection quantifiable	*** 10,000s of peptides per MS injection quantifiable	* 10s - 100s peptides per MS injection quantifiable
Selectivity/sensitivity/dynamic quantification range	** 4 orders of magnitude per MS injection	** 4 orders of magnitude per MS injection	*** 4–5 orders of magnitude per MS injection
Reproducibility/data consistency	*** High, due to peptide-centric scoring analysis	* Low, due to stochastic sampling in DDA	*** High, due to targeted data acquisition
Retrospective targeting (using chromatogram extraction)	*** Possible on MS1 and MS2 level	** Possible on MS1 level only	* Not possible due to targeted data acquisition

*Least optimal performance.
**Medium performance.
***Best performance.

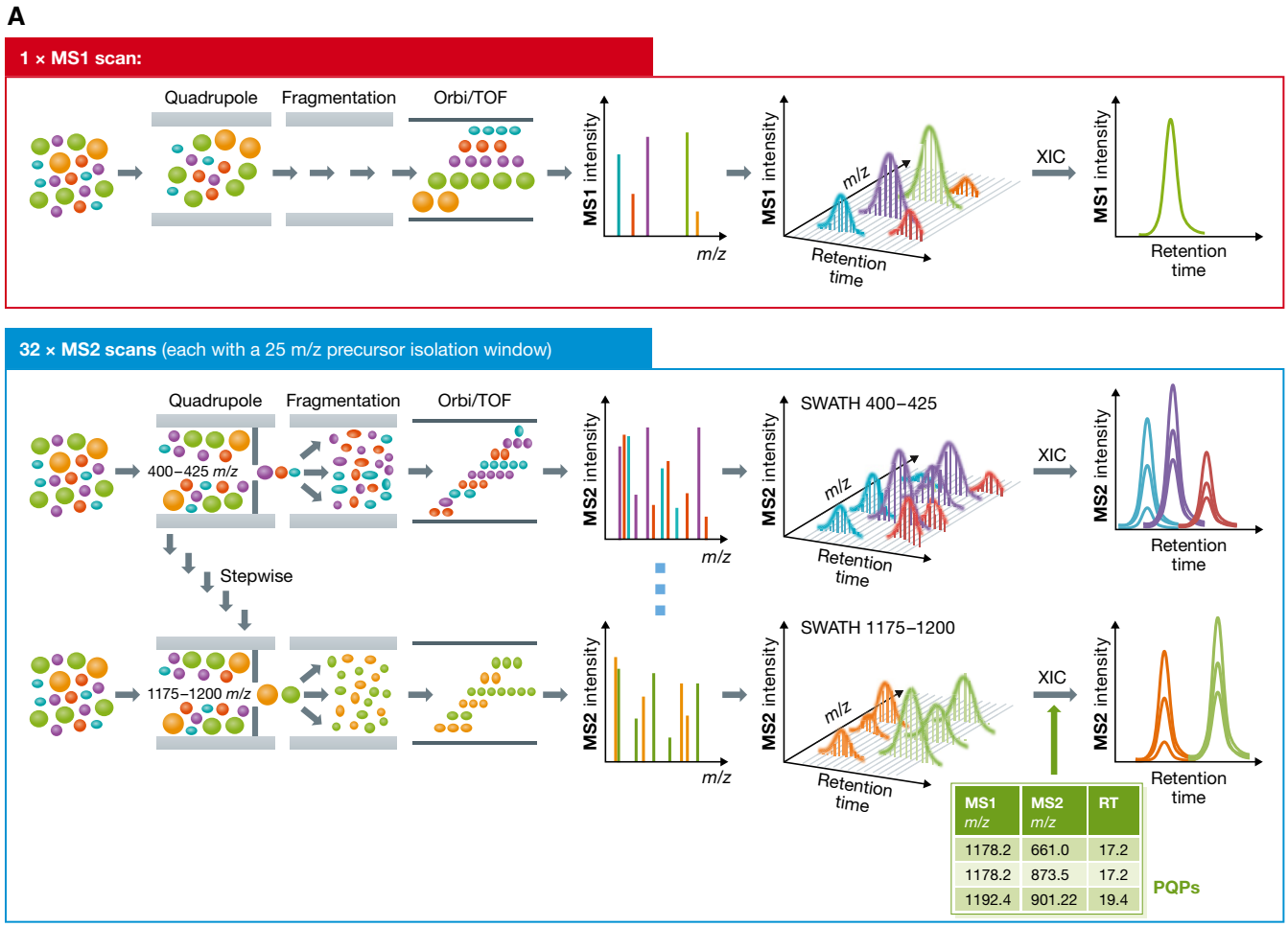
established. For an overview, see Appendix Fig S1, as well as recent reviews (Chapman *et al*, 2014; Bilbao *et al*, 2015). While the term "SWATH" became a registered trademark of SCIEX in the context of Q-TOF instrumentation, the company Biognosys trademarked the name "Hyper Reaction Monitoring" (HRM) for an analogous mode of data acquisition on Orbitrap instrumentation (Box 1). Throughout this tutorial, we only use the term "SWATH-MS", independently from the underlying instrument type, and the generic term "DIA" when referring to the breadth of all data-independent acquisition strategies.

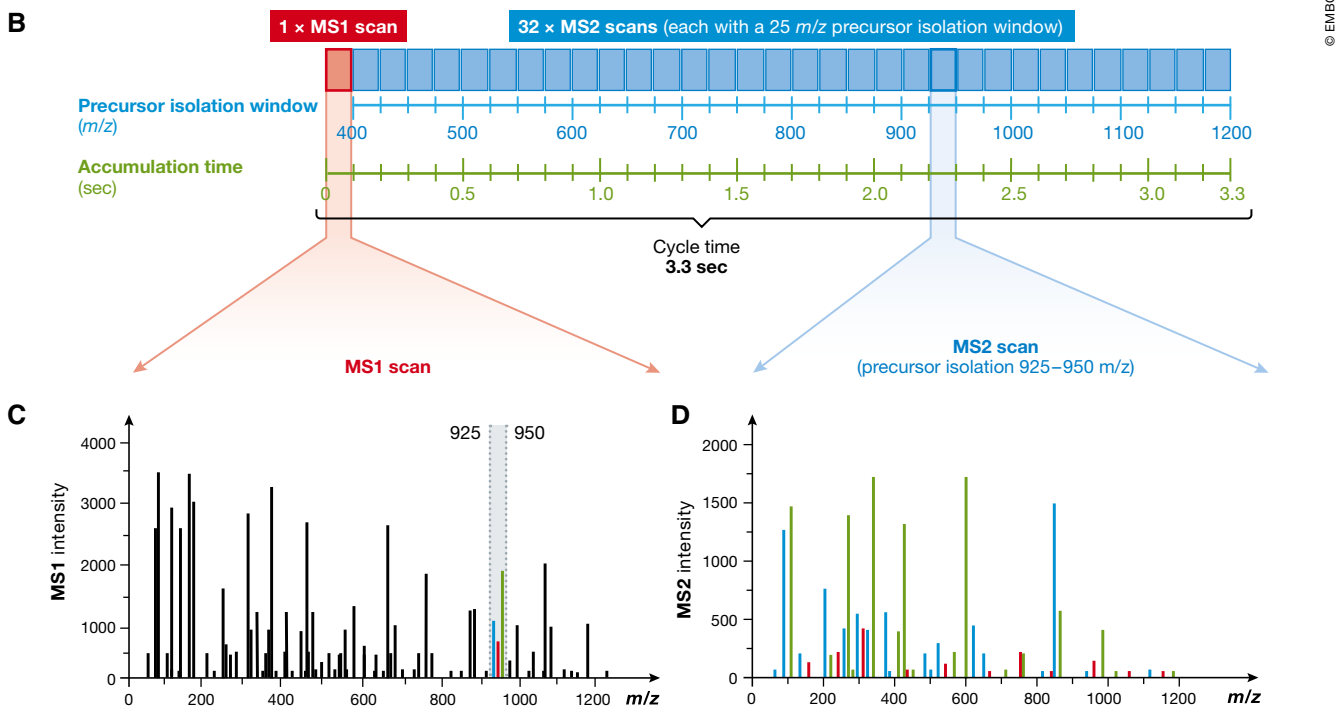
When is SWATH-MS the method of choice for my proteomic study?

The major advantage of SWATH-MS is that it supports quantitative analyses of peptides covering 1,000s of proteins with a high quantitative consistency and accuracy. It is ideally suited for projects that entail a large number of samples and that require accurate and reproducible quantification for the major fraction of the expressed proteome or peptidome in each sample. Typical projects that require exactly these properties include for example biomarker studies (Liu *et al*, 2014; Muntel *et al*, 2015; Kulkarni *et al*, 2016; Ortea *et al*, 2016), genetic association studies (Liu *et al*, 2015; Okada *et al*, 2016; Williams *et al*, 2016), clinical drug/perturbation studies (preprint: Litichevskiy *et al*, 2018; Tan *et al*, 2017; Keam *et al*, 2018) or exploratory basic research (Collins *et al*, 2013; Lambert *et al*, 2013; Parker *et al*, 2015a; Schubert *et al*, 2015b). SWATH-MS is also particularly well suited for studies that need fast analyses using LC gradient lengths below 60 min (Vowinckel *et al*, 2018). Proteome coverages of ~50% of the MS-detectable proteome have been achieved in complex mammalian samples in a single-shot analysis (Bruderer *et al*, 2017; Kelstrup *et al*, 2018). A current drawback

Figure 1. Principle of sequentially windowed data-independent acquisition in SWATH-MS.
(A) SWATH-MS measurements are performed on fast scanning hybrid mass spectrometers, typically employing a quadrupole as first mass analyser and a TOF or Orbitrap as second mass analyser. In SWATH-MS mode, typically a single precursor ion (MS1) spectrum is recorded, followed by a series of fragment ion (MS2) spectra with wide precursor isolation windows (for example 25 *m/z*). Through repeated cycling of consecutive precursor isolation windows over a defined mass range, a comprehensive data set is recorded, which includes continuous information on all detectable fragment and precursor ions. Hence, extracted ion chromatograms can be generated on MS2 as well as MS1 level. For the analysis of SWATH-MS data, a peptide-centric scoring strategy can be employed, which requires prior knowledge about the chromatographic and mass spectrometric behaviour of all queried peptides in form of peptide query parameters (PQPs). (B) The SWATH-MS data acquisition scheme described by Gillet *et al* (2012) for a Q-TOF mass spectrometer uses 32 MS2 scans with defined increments of 25 *m/z*, starting at 400 *m/z* and ending at 1,200 *m/z*. One full MS1 scan is recorded at the beginning. By applying an acquisition time of 100 ms per scan, a total cycle time of ~3.3 s is achieved. (C) The MS1 full scan detects all peptide precursors eluting at a given time point. For example, in the mass range from 925 to 950 *m/z*, three co-eluting peptide species are detected (green, red and blue). (D) The corresponding MS2 scan with a precursor isolation window of 925–950 *m/z* represents a mixed MS2 spectrum with fragments of all three peptide species.

 © 2018 The Authors





© EMBO

Figure 1.

© 2018 The Authors

Box 1: Definitions of frequently used terms in the context of SWATH-MS

Term	Definition
DIA	Data-Independent Acquisition (DIA)—here, we use DIA as an umbrella term for mass spectrometric acquisition methods that continuously acquire fragment ion (MS2) spectra in an unbiased fashion, without requiring the detection of peptide precursor ions in an MS1 survey scan (as in DDA) nor prior knowledge about peptide precursor *m/z* values (as in SRM and PRM). Specific acquisition methods of the DIA family include for example SWATH-MS, Shotgun-CID, MS^E, XDIA, MSX, AIF FT-ARM and others
SWATH-MS	Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra (SWATH-MS)—represents a specific variant of DIA, performed on hybrid full-scan instruments (preferably Q-TOF and Q-Orbitrap). The term SWATH is a registered trademark of SCIEX. In SWATH-MS data acquisition, successive pre-defined ranges of precursor *m/z* values are isolated and subjected to co-fragmentation (Gillet *et al*, 2012). Peptide-centric scoring of SWATH-MS data can be performed for example by using prior knowledge in form of a spectral library
HRM	Hyper Reaction Monitoring (HRM)—Synonym for SWATH-MS. The term has been implemented in the context of data acquisition on Orbitrap mass analysers and is a registered trademark of Biognosys
Peptide-centric scoring	A data query strategy which starts with a pre-defined list of target peptides and tests whether those peptides are detectable in the data with a certain confidence. It can be applied to individual MS2 spectra or to extracted ion chromatograms (XICs). For a peptide-centric scoring analysis, peptide query parameters need to be readily available. Typically, data acquired in SRM or PRM mode are analysed by peptide-centric scoring, but also data acquired by DIA methods, such as SWATH-MS, can be analysed in this way. An equivalent term also used in the literature is "targeted data extraction"
Spectrum-centric scoring	A data analysis type which aims at finding the peptide sequence(s) from a user-specified proteome that explain(s) a given MS2 spectrum best. It is typically applied in the context of discovery-driven proteomics with data acquired by data-dependent acquisition (DDA), but also data acquired by different DIA methods can be analysed in this way.
Peptide query parameters (PQPs)	Compendium of parameters required for peptide identification by peptide-centric scoring. PQPs are stored in a table format and include (i) optimal (proteotypic) peptides to target for a given protein, (ii) chromatographic elution times of those peptides on the applied chromatography setup, (iii) most intense fragment ions (typically four to six) generated under the applied fragmentation conditions, (iv) charge state(s) of precursors and fragment ions and (v) relative ion intensity of all selected fragments. PQPs can be derived from previous discovery-driven experiments, from which all peptide identifications are summarized in form of a spectral library
(Targeted) Peptide assay	Synonym for peptide query parameters (PQPs). Term mainly used in the context of SRM, where different tier levels (1–3) of analytical assay validation have been defined (Carr *et al*, 2014)
Spectral library	Compendium of MS2 spectra confidently assigned to a specific peptide sequence, typically acquired by discovery-driven proteomics using data-dependent acquisition. In case, several MS2 spectra refer to the same peptide sequence either the best scoring spectrum or an average consensus spectrum gets reported. Peptide retention time information can also be stored in a spectral library file and normalized retention times can be generated through retention time re-alignment using reference peptides. Alternatively, spectral libraries can be generated from deconvoluted pseudo-MS/MS spectra directly from DIA data (DIA-Umpire; Tsou *et al*, 2015)
Transition	Pair of a precursor and one corresponding fragment ion *m/z* value. Mainly used in the context of SRM
Targeted proteomics	Umbrella term for mass spectrometric methods that aim at quantifying a list of pre-defined proteins, peptides or PTM-peptides of interest. In the two classical targeted proteomic approaches, SRM and PRM, the data acquisition itself is performed in a targeted fashion. However, also data acquired "untargeted", using for example DIA measurements, can be analysed using a peptide-centric data analysis strategy, which classifies SWATH-MS as targeted proteomic approaches

of SWATH-MS compared to the classical targeted proteomic approaches (SRM or PRM) is that peptide quantification with SWATH-MS is still three- to 10-fold less sensitive (Gillet *et al*, 2012; Liu *et al*, 2013; Schmidlin *et al*, 2016). Hence, targeted data acquisition remains the better option for projects that involve quantification of particularly low-abundant proteins and peptides with maximal accuracy. A further drawback of SWATH-MS in comparison with DDA-based methods is the required upfront effort on experimental or *in silico* spectral library and PQP generation and optimization (Table 1).

In the context of very large-scale quantitative proteomic analyses, two alternative mass spectrometric strategies are currently used successfully in the field in addition to SWATH-MS. The first is the classical label-free DDA proteomics workflow, where quantification is based on precursor ion (MS1) intensities or spectral counts, and which can possibly be combined with peptide fractionation techniques to improve proteome coverage (Lawrence *et al*, 2015; Geyer *et al*, 2016; Frejno *et al*, 2017). An important improvement for the application of MS1 quantification to DDA data sets was the development of analysis tools that allow the transfer of peptide identifications between samples and thereby improve the completeness of the quantitative data matrix (Prakash *et al*, 2006; Mueller *et al*, 2007; Cox *et al*, 2014). However, even when using these tools, the number of missing values in DDA data sets still remains higher than for data acquired in SWATH-MS mode, especially for peptides and proteins in the low concentration range (Bruderer *et al*, 2015, 2017; Kelstrup *et al*, 2018). While direct comparisons of an optimal label-free MS1/DDA workflow versus a SWATH-MS workflow are challenging, several papers have demonstrated that when the same sample is injected under the same conditions using the same mass

 © 2018 The Authors

spectrometer operated once in DDA and once in SWATH-MS mode, SWATH-MS outperforms DDA in terms of detectable peptides and associated proteins as well as measurement reproducibility (Bruderer *et al*, 2015; Kelstrup *et al*, 2018).

A second popular strategy for large-scale quantitative proteomics relies on isobaric labelling, using, for example, tandem mass tags (TMT) (Thompson *et al*, 2003) or isobaric tags for relative and absolute quantitation (iTRAQ) (Unwin *et al*, 2005). Frequently, isobaric labelling is followed by an extensive peptide fractionation procedure. The resulting fractions are then analysed individually by DDA mass spectrometry (Chick *et al*, 2016; Roumeliotis *et al*, 2017). With state-of-the-art TMT reagents 10 (McAlister *et al*, 2012) or 11, samples can be mixed and analysed simultaneously, leading to minimal sample preparation biases and highly consistent and complete data matrices within a set of multiplexed samples. While early implementations of this method suffered from quantitative ratio compression due to interferences in reporter ions from co-eluting and co-fragmenting peptides, this has been addressed to some extent by using optimized data acquisition and analysis methods (Ting *et al*, 2011; Savitski *et al*, 2013; McAlister *et al*, 2014; Ahrne *et al*, 2016; O'Brien *et al*, 2018; Sonnett *et al*, 2018). Labelling samples with isobaric tags can be an optimal workflow for comparative analysis of medium-sized projects; however, if 100s of samples need to be analysed, the issues of data incompleteness and batch effects can become apparent again across sets of multiplexed samples. To date, studies that directly compare isobaric tagging with SWATH-MS are missing.

The overall intent of this tutorial is to guide readers towards performing their own SWATH-MS measurements. We give guidelines on how to set up and plan a SWATH-MS experiment, how to perform the mass spectrometric measurement using data-independent acquisition and how to analyse SWATH-MS data using peptide-centric scoring. Furthermore, concepts on how to improve SWATH-MS data acquisition, potential trade-offs of parameter settings and alternative data analysis strategies are discussed.

Setting up and planning a SWATH-MS experiment

If a SWATH-MS study progresses towards the analysis of 100s and eventually 1,000s of samples, particular attention should be paid to the feasibility of producing comparable data of good quality, both longitudinally on a single instrument, as well as across multiple instruments of the same type, and conceivably across different instrument platforms. Encouraging progress has been made demonstrating the comparability of SWATH-MS data generated between laboratories using standard samples (Collins *et al*, 2017).

Cumulative instrument contamination during a measurement series, caused for example by contaminants such as lipids, polymers or detergents, is a major concern and extra care must be taken to produce samples that are mostly devoid of such contaminants. Contaminants are an especially important issue in the context of SWATH-MS, because we observed faster and more severe instrument performance loss (such as charging effects and sensitivity issues) in SWATH-MS mode than in DDA or PRM mode on the same instrument. One possible explanation for this observation is that the instrument operated in SWATH-MS mode has a substantially higher ion flux in the fragment ion scans (which compose > 90% of the data acquisition time), which in turn means that also more sample contaminants and impurities might enter and contaminate the instrument, leading to a faster performance loss. Therefore, monitoring the performance of the mass spectrometer and maintaining it at an acceptable level are an important prerequisite. Recent efforts have been undertaken to develop software tools that enable systematic tracking of instrument performance and that can be applied also to SWATH-MS data (Rudnick *et al*, 2010; Wang *et al*, 2014; Bereman *et al*, 2016; Chiva *et al*, 2018).

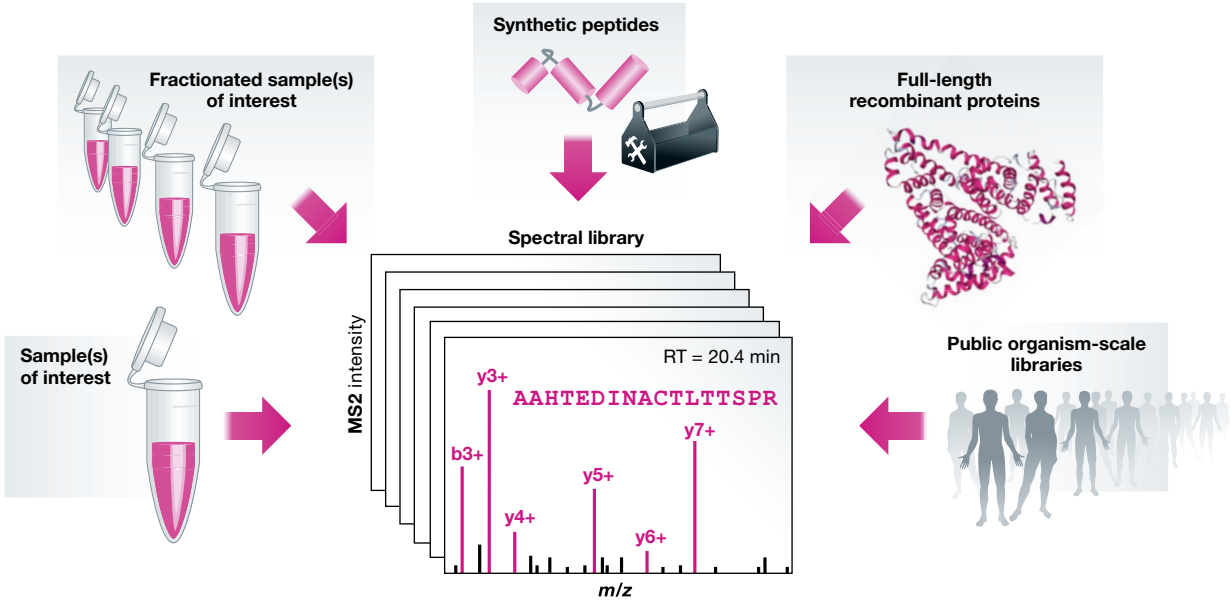
When planning a label-free large-scale proteomic experiment, statistical considerations for the experimental design, such as group size, biological and technical variability or achievable sensitivity and selectivity, should be taken into account (Krzywinski & Altman, 2014b). We suggest that particular attention should be paid to proper randomization and blocking (Krzywinski & Altman, 2014a) of a sufficient number of biological and technical sample replicates (Blainey *et al*, 2014) to ensure optimal statistical power (Krzywinski & Altman, 2013) during the downstream data analysis process.

Prior knowledge required for peptide-centric scoring and spectral libraries

The underlying concept behind SWATH-MS is that empirically derived prior knowledge of the mass spectrometric and chromatographic behaviour of peptides of interest can be used to selectively extract peptide-specific information from highly convoluted SWATH data in a targeted fashion (Gillet *et al*, 2012). This required prior knowledge is referred to as "peptide query parameters" (PQPs). It is worth noting that "prior" in this context indicates that PQPs should be available as a prerequisite before data analysis is undertaken, while the actual acquisition of the SWATH-MS data itself does not depend on the availability of PQPs.

What kind of information do PQPs contain? In detail, the information includes (i) the peptide sequence(s) to monitor for a given protein, (ii) the dominant precursor ion m/z value(s) of the peptide(s) and thus the charge state distribution, (iii) the four to six most intense fragment ion m/z values for the peptide(s) under the applied fragmentation conditions, (iv) information about the expected fragmentation pattern under the applied conditions, i.e. the relative fragment ion intensities, and v) the expected retention time of the peptide(s) and thus the associated fragment ion signals, ideally normalized to a reference. PQPs can commonly be obtained from a spectral library [or potentially chromatogram library (Sharma *et al*, 2014)] and are stored in a table format as shown in Fig 2. Computational pipelines integrating all steps of spectral library generation and PQP extraction have been developed that simplify and standardize this process and are available for example within Skyline (Egertson *et al*, 2015), PeakView (SCIEX), Spectronaut Pulsar (further referred to as Spectronaut) (Bruderer *et al*, 2015) (Biognosys) and the Trans-Proteomic Pipeline (TPP) (Deutsch *et al*, 2010). Further tools to prepare and convert [specL (Panse *et al*, 2015), Fraggle/Tramler/Franklin (Teleman *et al*, 2017)] or extend [SwathXtend (Wu *et al*, 2016)] sets of peptide query parameters are also available. Particularly large spectral libraries can be optimized and constrained by using MSPLIT-DIA (Wang *et al*, 2015).

© 2018 The Authors



Peptide query parameters (PQPs) for target peptide AAHTEDINACTLTTSPR

Protein	Peptide sequence	Precursor ion m/z	Fragment ion m/z	Precursor charge	Fragment charge	Fragment ion type	Fragment relative intensity	Normalized retention time
Raf1	AAHTEDINACTLTTSPR	929.4416	775.4308	2	1	y7	70	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	674.3832	2	1	y6	10	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	561.2991	2	1	y5	40	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	460.2514	2	1	y4	20	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	359.2037	2	1	y3	100	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	280.1404	2	1	b3	60	45.3

© EMBO

Figure 2. What are peptide query parameters (PQPs) and where do these parameters come from?
PQPs contain information about the chromatographic and mass spectrometric behaviour of a given peptide, as exemplified here for the peptide AAHTEDINACTLTTSPR. Various different input sample types can be used for the purpose of PQP generation. Typically, those samples are analysed in initial DDA measurements, and the results are summarized in the form of one or several spectral library files. From the spectral library file(s), the relevant PQPs are extracted by filtering the identified peptide coordinates using the indicated criteria. PQPs contain information about: the underlying protein, peptide sequence, precursor *m/z*, fragment *m/z*, precursor and fragment charge, fragment ion type, expected relative fragment ion intensities and normalized retention time (retention time relative to a set of reference peptides, iRT).

For further details on how to create spectral libraries and PQPs from DDA data, we refer to a recent protocol paper (Schubert *et al*, 2015a). In the following paragraphs, we will discuss the types and sources of information that can be used to generate spectral libraries from which PQPs can be extracted.

An alternative to acquiring DDA runs for spectral library generation represents spectrum-centric scoring of the DIA data. While such algorithms have been proposed early in the development of DIA acquisition schemes, recently developed algorithms such as DIA-Umpire (Tsou *et al*, 2015), Group-DIA (Li *et al*, 2015) and Spectronaut Pulsar (Biognosys) make specifically use of the improved data quality of the latest generation instruments. Generally, these algorithms generate a cumulative spectral library of a related set of samples and achieve similar coverage as DDA runs of unfractionated samples.

Representative sample types for spectral library generation

Here, one has to choose between endogenous sources of peptides from the samples of interest, synthetic analogues of those peptides or recombinant full-length proteins (Fig 2). The majority of

 © 2018 The Authors

SWATH-MS studies to date have used a side-by-side characterization of the samples of interest by DDA analysis for generation of spectral libraries. These libraries frequently include a sample fractionation step prior to DDA analysis, which can be beneficial from a sensitivity perspective for post-acquisition peptide queries (Rosenberger *et al*, 2014; Zi *et al*, 2014; Selevsek *et al*, 2015). This is because the sensitivity and coverage of single-shot DDA analysis are frequently lower than that of SWATH-MS data (Bruderer *et al*, 2015; Kelstrup *et al*, 2018). Therefore, the SWATH-MS data would not be fully covered by a "single shot" DDA spectral library generation strategy. The strategy focusing on repeated DDA analysis of non-fractionated samples is sometimes favoured because it is straightforward to implement; however, primarily for library completeness and quality reasons, other sources of peptides can be considered.

Chemically synthesized peptides have long been proposed as a source of prior knowledge, in particular, with respect to the development of SRM assays (Kuster *et al*, 2005; Picotti *et al*, 2010), and large-scale efforts to synthesize and measure peptides for several organisms on a proteome-wide scale have been reported (Picotti *et al*, 2013; Schubert *et al*, 2013; Kusebauch *et al*, 2016; Zolg *et al*, 2017). This approach has several advantages: (i) all proteins can be represented in the library, irrespective of whether they have been previously observed, (ii) high-quality MS2 spectra and derived PQPs can be generated, because the synthetic peptides can be analysed at very high concentrations and represent ground truth, and (iii) the error rate in the spectral library generated should be close to zero. In some cases, the selection of which peptides to synthesize for a given protein has been driven by prior empirical observation. However, since there has been so far no species with absolute complete proteome coverage by DDA methods, sets of proteotypic peptides for each protein have also been computationally predicted (Mallick *et al*, 2007). Since computational prediction methods have turned out to be less reliable than expected (Searle *et al*, 2015), empirical peptide selection is still preferred over computational prediction. More recent efforts to create synthetic proteomes for general purposes in proteomics have extended to a very large scale (> 330,000 human peptides) by synthesizing much larger numbers of peptides per protein and by also including PTMs and common sequence variants (www.proteometools.org; Zolg *et al*, 2017). A useful extension of this approach is to create full-length proteins by recombinant methods or *in vitro* transcription/translation systems. In this way, the most suitable peptides per protein for analysis can be determined empirically (Stergachis *et al*, 2011; Matsumoto *et al*, 2017). While those large-scale synthetic peptide and protein MS resources have not yet been fully exploited for SWATH-MS analysis, it seems likely that they will be useful resources going forward. Hybrid libraries consisting of endogenous samples and synthetic peptides to increase coverage may also be an attractive option and have proved useful in the case of an organism-scale library for *Mycobacterium tuberculosis* (Schubert *et al*, 2015b).

Comprehensive versus sample-specific spectral libraries

A natural extension of the ideas discussed above is to attempt to characterize peptides by DDA from all proteins in a given species to generate a comprehensive "off the shelf", species-specific library for general use, obviating the need to generate experiment- or sample-specific libraries. These libraries could include data from endogenous samples and/or synthetic proteomes. Presently, such organism-scale spectral libraries are publicly available for *S. pyogenes* (Karlsson *et al*, 2012), *S. cerevisiae* (Picotti *et al*, 2013; Selevsek *et al*, 2015), *M. tuberculosis* (Schubert *et al*, 2013, 2015b), *H. sapiens* (Rosenberger *et al*, 2014; Kusebauch *et al*, 2016; Zolg *et al*, 2017), phyllosphere-colonizing bacteria (*M. exotorquens*, *P. syringae* and *S. melonis*) (Muller *et al*, 2016) and *E. coli* (Ludwig *et al*, in preparation). Further, the number of public repositories for various types of MS data that could conceivably also be used as prior knowledge in the analysis of SWATH-MS data continues to increase (Craig *et al*, 2004; Martens *et al*, 2005; Deutsch *et al*, 2008; Picotti *et al*, 2008; Sharma *et al*, 2014; Whiteaker *et al*, 2014; Wilhelm *et al*, 2014; Zolg *et al*, 2017). These include SWATHAtlas (www.swathatlas.org), which focuses on SWATH-MS and provides spectral libraries in formats directly compatible with most software tools for peptide-centric analysis of SWATH-MS data.

An important consideration for the use of such public libraries will be portability of information between instrument types and between laboratories. For example, efforts have been made to evaluate the effect of using DDA spectra generated on different instrument types as prior knowledge (Toprak *et al*, 2014). Although the best comparability is achieved when fragment ion spectra are generated on the same instrument platform, the data also indicated that fragment ion spectra acquired from instruments using "beam type" collision-induced dissociation (e.g. QqQ, QqTOF and Orbitrap operated in HCD mode) produce spectra that are sufficiently comparable for their effective use as prior knowledge in peptide-centric analysis (de Graaf *et al*, 2011; Zolg *et al*, 2017). The portability of chromatographic information between laboratories has been advanced through the use of normalized retention times (Norbeck *et al*, 2005) such as indexed retention time (iRT) (Escher *et al*, 2012). The accuracy of such methods is generally tolerant to changes in gradient length and column dimensions; however, larger errors will result from changes in other factors that may significantly affect the peptide elution order such as mobile/stationary phases or column temperature. An inter-laboratory evaluation study recently showed that an organism-scale spectral library could effectively be used to analyse SWATH-MS data generated from 11 different laboratories worldwide (Collins *et al*, 2017).

Another important issue to consider when working with very large-scale spectral library resources and querying very high numbers of peptides in many samples measured by SWATH-MS is appropriate error rate control. In such analyses, it is common that a significant fraction of queried peptides from the library are actually not present in the samples of interest at a detectable level and, as such, represent channels in which false positives may appear and accumulate if many runs are analysed. If SWATH extraction results are summarized at the inferred protein level, the error rate can be further inflated by such false positives (Reiter *et al*, 2009; Rosenberger *et al*, 2017a). Therefore, it is important to control false discovery rates at the protein level, rather or in addition to the peptide level and to compute false discovery rates globally for a complete experiment rather than at the per-file level. Rosenberger *et al* (2017a) recently established statistical concepts derived from discovery proteomics that can be applied to appropriately control the error rate of detected peptides and inferred proteins in SWATH-MS.

In conclusion, using organism-scale "off the shelf" libraries might be convenient from the perspective of reducing the effort of creating project specific libraries, but their use may eventually result

© 2018 The Authors

in some loss of statistical power due to the additional multiple testing burden and necessary correction that comes from querying such a large number of peptides. Alternatively, researchers need not necessarily query all of the peptides in a spectral library, but can also focus on just a subset of proteins that are of interest for their specific biological question.

Which and how many peptides should be queried for?—From targeted to discovery uses of SWATH-MS

In the initial implementation of SWATH-MS, the primary goal was to use peptide-centric scoring based on prior knowledge combined with sequentially windowed data-independent acquisition (Gillet *et al*, 2012) to develop an approach that generates data resembling those from classical targeted proteomics approaches such as SRM but at a largely extended scale of analytes. Hence, a discrete set of peptides from proteins related to a particular biological question was targeted for quantification. However, it was soon realized that SWATH-MS data sets offer the opportunity to characterize complex samples in a more comprehensive manner by extracting and scoring ion chromatograms for a maximum number of peptides, by using sample-specific or combined libraries. Over time, the number of peptides that were queried in comprehensive SWATH-MS studies has grown significantly to the point that, with the appropriate error-rate control (Rosenberger *et al*, 2017a), it is now possible and reasonable to query for tens to hundreds of thousands of peptides derived from organism-scale spectral libraries in a fashion that resembles discovery-based proteomics much more than targeted proteomics in its scope (Bruderer *et al*, 2017).

The choice of whether to choose a targeted approach and extract peptides for only a smaller set of proteins of interest, or whether to extract data for all peptides contained in a large-scale spectral library, will depend on the specificity of the hypothesis that underlies the targeted approach, i.e. how useful the set of proteins being measured is for answering a given biological question. If a biological hypothesis can be addressed with a selected set of target proteins, then increasing the number of queries by including peptides from a large-scale library will increase the multiple testing burden. If the hypothesis cannot be addressed by few selected proteins and requires a data-driven investigation, such as exemplified by classification problems, then extracting a maximum number of peptides from a large-scale library will likely provide the most informative result, despite the increased number of statistical tests that need to be performed.

The possibility that specific peptide queries might result in false-positive signals arising from contaminating peptide species that are similar to the target peptide in their mass spectrometric and chromatographic parameter space has been a concern of peptide-centric scoring. This might be the case for peptides with a sequence that is similar to that of the target peptide, or for peptides for which the precursor mass difference between target and mismatch can be explained by a modification of the common backbone sequence. A recent study comparing standard database searching of DDA data with open modification searches of the same data found that this scenario occurs rather infrequently (Kong *et al*, 2017). The open modification search assigned a given fragment ion spectrum to a modified peptide sequence rather than to the unmodified peptide sequence originally assigned by the standard database search only for ca. 1% of the identified peptides. This suggests that false-positive identifications in SWATH-MS data due to unanticipated modified peptides might also occur rather infrequently, but nevertheless, close attention in ongoing algorithmic development should be paid to the likelihood of matching-related peptide species (Rost *et al*, 2012). Data analysis methods have now been developed that address the problem of related "peptidoforms" in SWATH datasets (Rosenberger *et al*, 2017b).

Overall, the high degree of fragment ion data completeness, both in the mass and chromatographic dimension, constitutes a main strength of data acquired by SWATH-MS and related DIA approaches, as it offers the opportunity of applying diverse and flexible data analysis strategies. It also equally supports different experimental strategies, including targeted quantification of selected analytes and discovery-driven studies of all detectable analytes, provided that the types of possible scoring errors are recognized and appropriately controlled.

Performing a SWATH-MS measurement

Setting up liquid chromatography for SWATH-MS

The LC setup used for SWATH-MS measurements does not differ from that used for targeted or DDA setups. Typically, peptide mixtures of 0.5–2 μg total peptide mass are separated by injection onto a nano-HPLC system equipped with a 15–50-cm-long column packed with reversed phase particles (e.g. 3 or 1.8 μm C18 material), operated at 300 nl/min flow rate and eluted with a solvent gradient of increasing organic composition. Note that ~5 μl/min microflow HPLC systems are also compatible with SWATH-MS, though at the cost of a three- to fivefold loss of sensitivity that can be compensated by injecting three to five times more peptide amount on column (Bruderer *et al*, 2017; Vowinckel *et al*, 2018).

Peptide-centric analysis of SWATH-MS data depends on a sufficient number of data points (typically in the range of 10) collected over the elution profile of a peptide peak to allow the accurate reconstruction of the respective chromatographic peak (Fig 3C). The 2–4 s cycle time required by the current windowed SWATH-MS methods to cycle through the precursor isolation range, together with the required number of measurements over the peak indicates that the average chromatographic peak width in SWATH-MS measurements should not fall below 20–40 s. This is in contrast to single-window DIA methods, which only cycle between a single MS1 and a single MS2 scan and which can accommodate the 2–3 s-wide chromatographic peaks produced by higher performance LC setups [for example AIF (Geiger *et al*, 2010) or MS^E (Silva *et al*, 2006)]. As MS instruments improve in scanning speed, narrower chromatographic peaks may increasingly become compatible with windowed SWATH-MS methods.

Interestingly, SWATH-MS copes well with shortening of the HPLC gradient length. While the number of acquired MS2 spectra and the number of peptides identified decrease proportionally with the gradient length in DDA mode, the deterministic nature of the MS2 sampling in SWATH-MS mode attenuates the attrition in number of peptides identified for shorter separation gradients (Fig 3A). The loss in the number of peptide identifications at shorter gradients in SWATH-MS mode seems, therefore, to be related to the increase in precursor co-elution and ensuing potential ion suppression and co-fragmentation effects rather than to

 © 2018 The Authors

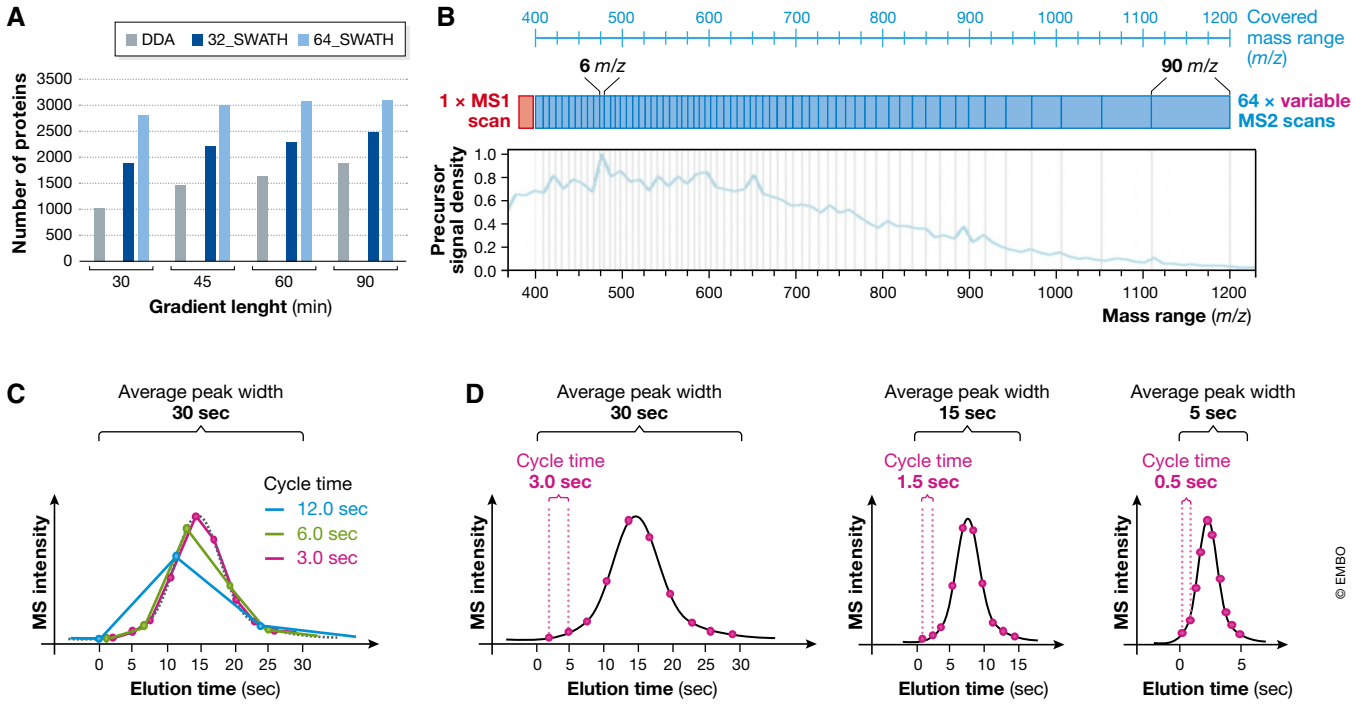


Figure 3. Setting up the optimal SWATH-MS data acquisition scheme.
(A) Effect of liquid chromatography gradient length on the number of identified proteins for technical triplicate injections of a trypsin-digested HEK cell lysate acquired either in DDA (grey bars), SWATH 32 fixed windows (dark blue bars) or SWATH 64 variable windows (light blue bars) on a Q-TOF instrument. (B) To improve precursor selectivity in SWATH-MS, an acquisition scheme using variable precursor isolation window widths can be used to partition the precursor density equally across all isolation windows. (C) Ten recorded data points are considered necessary for accurate reconstruction of a chromatographic peak. For a peak width of 30 s a cycle time of 3 s leads to 10 recorded data points (red), which allows appropriate reconstruction of the actual peak shape (grey dashed line). Longer cycle times, for example 6 s (green) or 12 s (blue), lead to under sampling and the correct peak shape can no longer be optimally reconstructed. (D) If the average peak width is reduced from 30 (left panel) to 15 (middle panel) or even 5 s (right panel), the cycle time needs to be decreased accordingly from 3 to 1.5 and 0.5 s, in order to maintain 10 data points over the elution profile.

undersampling. As MS instrumentation continues to improve, it is likely that SWATH-MS will reach higher selectivity and identification rates under short gradient regimens compatible with relatively high sample throughput.

In summary, we advise a 2-h nano-HPLC gradient for the acquisition of high-quality spectral libraries (Schubert *et al*, 2015a), while the use of shorter nano- (or micro-) HPLC gradients to acquire SWATH-MS data, for example in the range of 30–60 min, still provides results with good selectivity and proteome coverage at a significantly higher sample throughput.

Setting up the SWATH-MS data acquisition method

Finding an optimally balanced method

The optimal setup of precursor ion (MS1) and fragment ion (MS2) scans in a SWATH-MS acquisition scheme depends on a range of considerations, including the expected mass range for analytes of interest, underlying sample complexity, available mass spectrometer type with its specific resolution and scanning speed, the LC setup and its expected average chromatographic peak width, as well as the desired measurement selectivity and sensitivity. In practice, the following instrument parameters need to be considered: (i) the precursor *m/z* range to cover, (ii) the widths and number of precursor isolation windows, (iii) the fragment and precursor ion accumulation/scanning time and resolution, (iv) the chromatographic cycle time and (v) the number of injections per sample. In the following section, each of these parameters is discussed in detail. The experience of the authors is primarily focussed on Q-TOF instrument configurations, and our recommendations in that respect are based on empirical knowledge, optimization and simulation (Rost *et al*, 2012). However, as the quadrupole-Orbitrap is increasingly popular for SWATH-MS analysis, we also refer to successfully published methods on that configuration. A detailed overview of acquisition parameters for Q-TOF as well as Q-Orbitrap instruments is provided in Table EV1. Those parameters can serve as a starting point for the generation of SWATH-MS data of good quality; however, further optimization might allow an increase in the level of quantifiable proteins and peptides. Of final note, optimal SWATH-MS parameters for Q-TOF as well as Q-Orbitrap instruments are still subject to adaptations and might change significantly in the future.

Precursor *m/z* range to cover The precursor *m/z* range covered in a SWATH-MS measurement is defined by the adjacent set of precursor isolation windows that the instrument cycles through across the chromatographic separation. Ideally, this mass range should cover as completely as possible the *m/z* space of the proteins or peptides of interest. Analytes outside the specified range are not monitored

© 2018 The Authors

and can therefore neither be detected nor quantified. If the aim is to analyse a significant portion of the proteome and if trypsin is used as proteolytic enzyme during the sample preparation, the most peptide-rich region typically spans 400–1,200 *m/z* (Gillet *et al*, 2012). The range can be further reduced to 500–900 *m/z* (Egertson *et al*, 2015) depending on instrument and analyte constraints with a minimal attrition of quantified proteins. If other proteases than trypsin are used, or if SWATH-MS is applied to analytes other than peptides (e.g. metabolites), the precursor *m/z* range needs to be adapted accordingly. Coordinates for *m/z* ranges successfully used in selected publications on Q-TOF and Q-Orbitrap instrument are given in Table EV1.

Precursor isolation window placement The precursor isolation window width defines the range of peptide precursor masses that are co-isolated and co-fragmented in a given MS2 scan. Hence, the isolation window width directly influences the selectivity and the dynamic range of the measurement and in turn the sensitivity of peptide detection. It is probably the parameter that varies the most between various DIA methods, ranging from one isolation window of 800 *m/z* to hundred windows of 2 *m/z* (Appendix Fig S1).

Whereas the use of wider isolation windows allows the mass spectrometer to cycle faster through the predetermined precursor *m/z* range, it results in a higher number of peptide precursors being co-fragmented, more strongly convoluted MS2 spectra and lower sensitivity due to a limited intra-scan dynamic range. Conversely, using narrower isolation windows reduces the number of co-fragmented precursors and signal interference, but limits other parameters such as the covered precursor *m/z* range per injection or the number of data points recorded over the chromatographic peak profile.

Note that the precursor isolation window width does not need to be identical for all MS2 scans. Rather, variable window widths (Zhang *et al*, 2015) can be used to optimize the number or intensity distribution of precursor ion signals between isolation windows (Fig 3B). Such optimal variable window patterns can be applied on a per-sample basis or identically across multiple samples. Automated tools are available to facilitate such placement based on empirical data ("SWATH Variable Window Calculator" provided by SCIEX). In practice, we observed that the fine adjustment of the isolation window widths does not drastically influence the peptide identification or quantification results. For practical reasons, we therefore generated a "universal 64 variable window" acquisition scheme for TOF instruments (Table EV1) that seem compatible with most complex tryptic digests from various organisms ranging in complexity from bacteria to human. This precursor isolation scheme was devised to split ~50,000 experimental human tryptic precursor signals from a human cell line in 64 equal bins that vary from 5.9 *m/z* width for the range [472.4–478.3], where the density of precursor ion signals is the highest, up to 90.9 *m/z* width for the range [1,109.6–1,200.5]. The gain achieved by moving from 32 fixed to 64 variable windows has been demonstrated both qualitatively and quantitatively (Navarro *et al*, 2016). Bruderer *et al* optimized a SWATH-MS acquisition method for the Q-Exactive with 19 variable windows (Bruderer *et al*, 2015) and the Q-Exactive HF with 24 variable windows (Bruderer *et al*, 2017), whereas Kelstrup *et al* (Kelstrup *et al*, 2018) published a method with 70 windows for the latest and fastest Q-Exactive HF-X instrument (Table EV1).

Recently, alternative DIA acquisition methods have reported the use of deterministic but differential isolation windows during LC separation and a peptide's elution profile (Egertson *et al*, 2013; Moseley *et al*, 2018). However, their analysis requires specific deconvolution tools to decipher, post-acquisition, the contribution of the differential interferences arising during the peptide co-isolation process across the various isolation schemes.

Fragment and precursor ion accumulation time/resolving power For Q-TOF instruments, the MS acquisition or accumulation time defines how long the mass spectrometer accumulates ion signals for a given MS scan. With longer accumulation times, the signal-to-noise ratio for the acquired spectrum increases. Conversely, longer accumulation times increase the time the instrument requires to cycle through the series of MS scans. Therefore, the accumulation time should be chosen in conjunction with the number of precursor isolation windows (Fig 3C).

For Orbitrap mass analysers, the MS acquisition is subdivided into two parallelized time-constrained processes. First, the "injection time" is the time required to collect in the ion trap the desired number of charged species designated by the automatic gain control (AGC) parameter. Second, the "scan time" is the time required by the Orbitrap to record the mass spectrum depending on the set resolution. On new generation instruments (Q-Exactive and Fusion/Lumos), these two processes are parallelized, and thus, the more time-consuming step determines the cycle time. Typically, the maximal injection time is set according to the required scan time. Recent publications have reported methods with 19 (Bruderer *et al*, 2015) (Q-Exactive) and 24 variable precursor isolation windows (Bruderer *et al*, 2017) (Q-Exactive HF) at a resolution of 30,000 (at Full Width Half Maximum) or 70 windows (Kelstrup *et al*, 2018) (Q-Exactive HF-X) with fixed width of 9 *m/z* at a resolution of 15,000 (Table EV1). With the rapidly evolving specifications of these instruments, further optimization of the acquisition settings can be expected.

In summary, for Q-TOF instruments coupled to a nano-LC setup that results in an average chromatographic peak width in the range of ca. 30 s, we recommend to use an MS2 accumulation time in the range of 50 ms for each of 64 variable MS2 acquisition windows. This adds up to a total cycle time of ~3.3 s. On the Q-Exactive HF instrument, the use of 24 variable windows at 30,000 resolution has been described recently, leading to a similar cycle time (Bruderer *et al*, 2017). However, due to the added complexity of balancing automatic gain control, scan time and duty cycle effects on Orbitrap instrumentation, further optimization for given instrument configurations and sample types should be considered. We also recommend the inclusion of a 250 ms MS1 scan on Q-TOF instruments before each cycle of MS2 scans. Similarly, the inclusion of one or several MS1 scans is recommended for Orbitrap instruments. Information from the MS1 scan is increasingly being used by recent DIA analysis software, either to confirm the precursor mass of the peptide of interest upon peptide-centric targeted chromatogram extraction or to aid in the deconvolution of co-eluting fragment ion traces with spectrum-centric analysis tools.

Chromatographic cycle time In SWATH-MS, the "cycle time", also referred to as "duty cycle" or "sampling rate", corresponds to the sum of the accumulation times set for the MS1 and MS2 scan series

 © 2018 The Authors

(plus any instrument overhead time which is usually negligible). Hence, the cycle time defines how often along the chromatographic elution profile of a peptide the same ion gets recorded. The cycle time is critical for any chromatography-based quantification approach. Generally, around 10 acquired data points are considered necessary to reconstruct chromatographic peak shapes and to perform accurate quantification (Fig 3C) (Matthews & Hayes, 1976; Lange *et al*, 2008). However, the available time to collect 10 data points throughout an average chromatographic peak width can vary substantially, depending on the specific chromatographic setup. In an experiment with an average peak width (at base) in the range of 30 s (typical peak width in a 3 μm C18 particles nano-HPLC setup with gradient length in the range of 1–2 h), a cycle time of 3 s is appropriate. However, if a higher resolution chromatographic system is employed, the average peptide elution peak width might be significantly shorter, for example 5 s. In such cases, the cycle time must be decreased, e.g. to 0.5 s, to ensure the constant recording of 10 data points over the elution profile (Fig 3D).

Number of injections per sample For large-scale proteomic studies that include 100s of samples, a single injection per sample is strongly preferable to maintain sample throughput and quantification consistency. However, for smaller scale projects, multiple injections per sample (N), each covering a different m/z range, might be an attractive alternative to improve selectivity and sensitivity of the SWATH-MS measurement (Panchaud *et al*, 2009; Ting *et al*, 2017). This allows the reduction in precursor isolation window width by a factor of N, while ensuring that a large precursor m/z range can be covered.

Setting the collision energy in SWATH-MS

Whereas for DDA, SRM and PRM acquisition methods, the collision energy can be set individually on a per-precursor basis (based on precursor m/z value and charge state), or even optimized for a specific transition (SRM), the collision energy in SWATH-MS can only be set on a per-isolation window basis. In other words, if two peptides with different charge states are selected in the same precursor isolation window, they will be fragmented with the same collision energy, which is likely to be suboptimal for at least one of the charge states. However, this suboptimal peptide fragmentation can be compensated for, to some degree, as long as the measured relative fragment ion intensities are similar to those in the spectral library used to query the data. This can be achieved by acquiring DDA fragment ion spectra for spectral library generation with a charge-state-independent collision energy equation, i.e. a collision energy varying with the precursor m/z but identical for all charge states. Note that this strategy may come at the cost of lost identifications in the DDA analysis due to suboptimal fragmentation of higher charge state peptides. This suboptimal fragmentation might be partially rescued by ramping the collision energy within a SWATH window, for example from −15 to +15 V, over the course of MS2 accumulation in both DDA and SWATH-MS mode.

In summary, for Q-TOF instruments, we recommend to set a collision energy proportional to the m/z of a 2+ charged precursor for DDA library generation and to apply the same collision energy of a theoretical 2+ precursor located in the centre of the precursor isolation window for SWATH-MS. Further, we recommend to systematically assess the fragmentation similarity between instruments and external public spectral libraries. Software tools that support such comparative analysis are readily available (Toprak *et al*, 2014). If significant differences between the fragmentation pattern of a spectral library and a SWATH-MS measurement are identified, it might be worth to optimize the used collision energy equation in the SWATH-MS measurements until a maximal similarity to the spectral library is found. This will ensure optimal portability of the peptide query parameters across instruments and laboratories.

Should adjacent precursor isolation windows be overlapping?

In the original SWATH-MS publication, we suggested to acquire consecutive SWATH windows with a 1 m/z precursor isolation m/z overlap (Fig 4A and Table EV1) (Gillet *et al*, 2012). This was implemented for two reasons: first, precursor ion isolation of the quadrupole mass analyser does not work with 100% efficiency over the whole mass range, but is compromised at the borders of the isolation windows (Fig 4B). Therefore, a small m/z overlap on both window edges compensates to some degree signal losses due to not perfectly square-shaped ion transmission efficiencies and ensures that precursors with m/z located at the border of the isolation window suffer as little as possible from signal attrition. Second, overlapping windows ensure maximal transfer of the complete isotopic pattern of precursors that would otherwise be split between two consecutive isolation windows for precursor masses close to a window edge (compare Fig 4C and D). Importantly, unless the entire precursor isotopic envelope gets isolated, the fragment ions will show a distorted isotopic pattern, which will potentially result in lower scoring of this peptide at data analysis stage (see section "Automated peak group scoring"). The SWATH-MS schemes published to date with 1 m/z window overlaps do not completely account for the problems mentioned above in the most extreme cases and, as such, slightly larger overlaps could be considered in future schemes. Note that, even if the precursor isotopic distribution is split between two consecutive isolation windows (as shown in Fig 4C), *label-free* quantification of that peptide using fragment ion intensities will still be accurate across runs, as long as the window geometries are reproducible and consistent between scans and injections (Egertson *et al*, 2015). A problem may occur, though, in the case of labelled SILAC-type experiments when one of the precursor isotopic distribution (e.g. the light form) would be split but the heavy is not. In this case, the light-to-heavy monoisotopic fragment ion ratios would be distorted due to the fact that part of the intensity of the monoisotopic peak at the fragment ion level is arising from the non-monoisotopic peaks in the precursor isotopic envelope.

Another possibility to minimize the impact of imperfect ion transmission efficiency is to purposely position the isolation window edges at masses where peptide m/z values are unlikely to occur as described by Egertson *et al* (2013). Such theoretically optimal window placement can be realized easily using the Skyline software (Egertson *et al*, 2015). Though this does not *per se* account for the signal attrition at the window borders, nor prevents precursor isotopes to split in different windows, it ensures that the monoisotopic precursor mass does not fall exactly on a window border.

In summary, the 1 m/z window overlapping suggested by Gillet *et al* (2012) causes a small loss in terms of selectivity, because the effectively measured windows are slightly broader. However, this is justified by positive effects like minimized signal losses for peptides

© 2018 The Authors

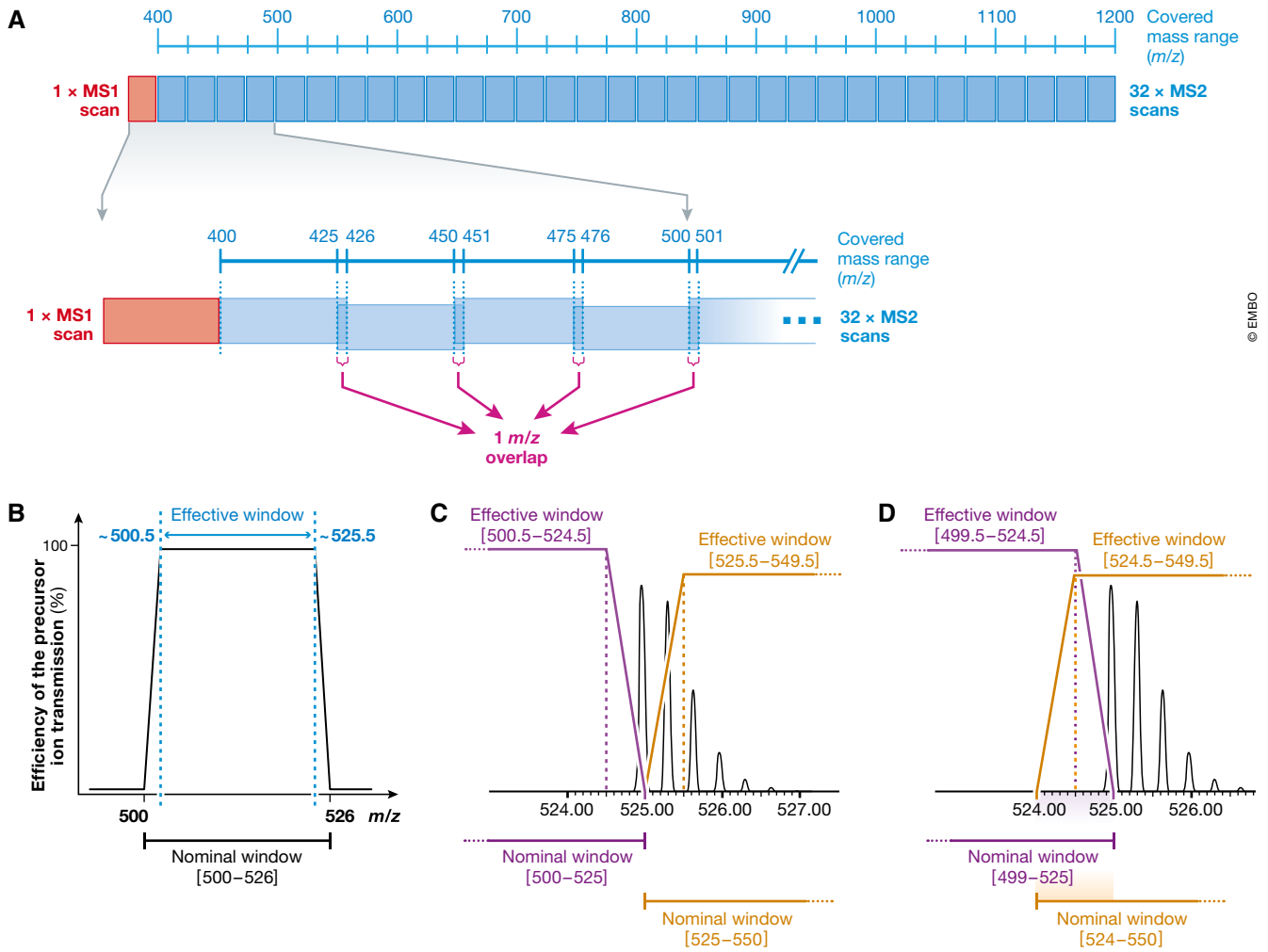


Figure 4. The rationale for using slightly overlapping precursor isolation windows.
(A) In the first implementation of SWATH-MS, a 1 *m/z* overlap between adjacent SWATH windows was used to compensate the inefficient ion transmission of the quadrupole at the edges of the precursor isolation window (B) and to limit the effect of precursor isotope splitting between windows (C and D). (B) Schematic representation of the efficiency of ion transmission with a state-of-the-art quadrupole mass analyser filtering for the [500–526] *m/z* range. (C) Theoretical isotopic distribution of the doubly charged peptide MLSYPITIGSLLHK (*m/z* = 524.965). If non-overlapping nominal windows are used [(500–525) and (525–550)], the isotopic profile is split between both windows and falls within the inefficient ion transmission range of the quadrupole [effective windows from ca. (500.5–524.5) and (525.5–549.5)]. Even if the window edge is placed at a mass where no precursor mass is supposed to occur (for example 525.1), the issue of inefficient ion transmission and loss of parts of the isotopic envelope would remain. (D) With overlapping nominal windows [(499–525) and (524–550)], most of the precursor isotopic pattern will be transmitted within the effective ion transmission range of the quadrupole [effective windows from ca. (499.5–524.5) and (524.5–549.5)].

falling at the sides of the isolation windows and more complete transfer of the isotopic envelope. If no window overlap is implemented, it is highly recommendable to perform optimal window placement as suggested by Egertson *et al* (2013).

Advanced DIA acquisition schemes with improved precursor selectivity
The selectivity of SWATH-MS is a direct function of the precursor isolation window width. Further, selectivity directly affects the achievable dynamic range and sensitivity of a SWATH-MS measurement. Recently, novel acquisition methods have been proposed to improve the selectivity of SWATH-MS (Fig 5).

Multiplexed DIA (MSX) In 2013, the MacCoss group reported the development of MSX (Egertson *et al*, 2013), a multiplexed DIA scheme. Instead of recording MS2 spectra for precursors selected in systematic SWATH windows of 20 *m/z* width, MSX records at each cycle fragment ions for peptides selected from different multiple *non-contiguous* isolation windows (e.g. five non-contiguous isolation windows of 4 *m/z*, Fig 5A). For the MSX setup, the overall absolute selectivity per spectrum is theoretically the same as for a 20 *m/z* wide SWATH-MS method. However, in MSX, a given peptide precursor will randomly co-fragment with different species in each cycle, resulting in different interferences over the elution profile of a peptide. In its current implementation, MSX requires an instrument with multiplexed trapping capabilities (ruling out Q-TOFs).

Offset isolation windows intra-run In the offset window approach, the precursor isolation window boundaries are shifted back and

 © 2018 The Authors

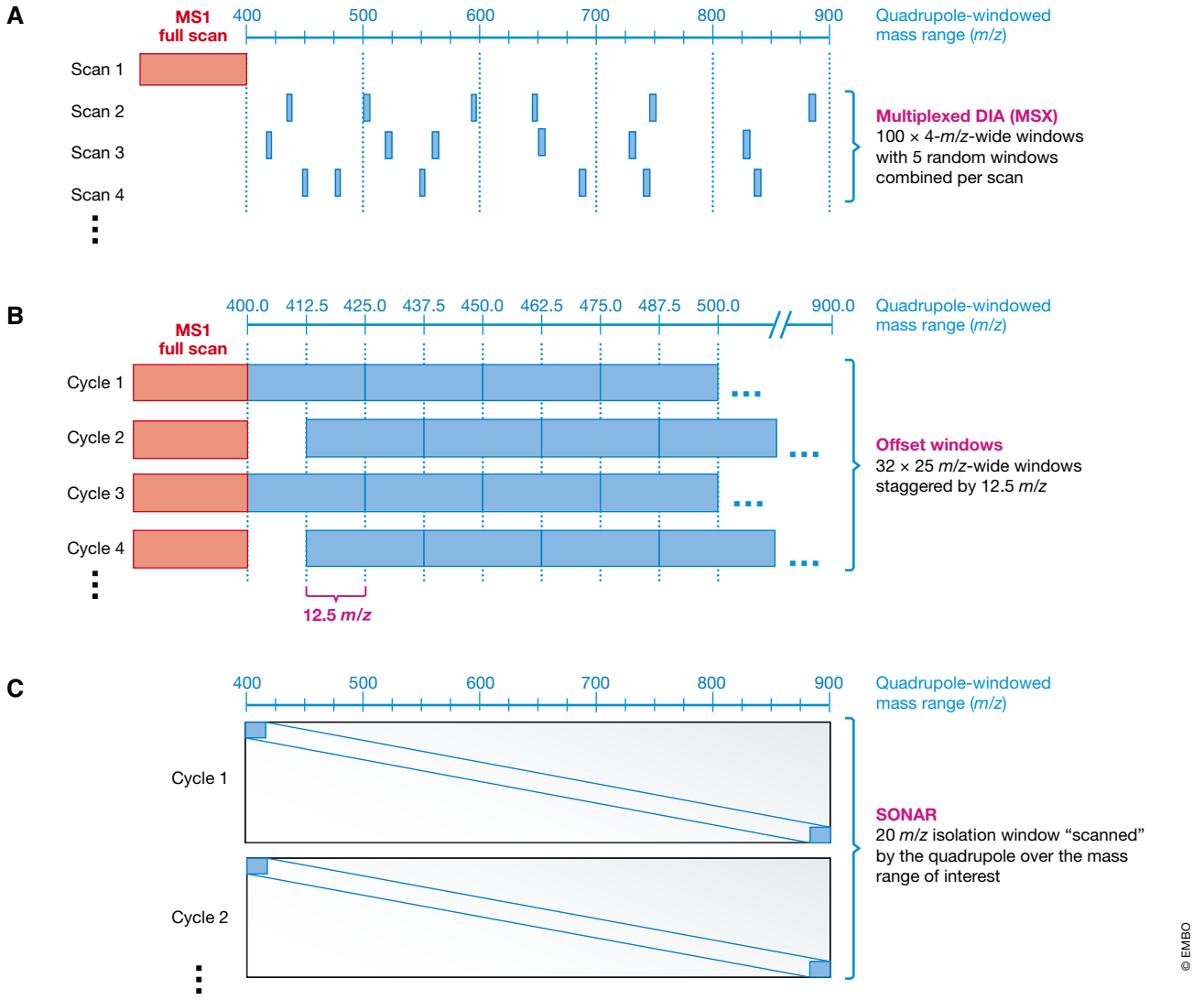


Figure 5. Advanced DIA acquisition schemes with improved precursor selectivity.
(A) Multiplexed DIA (MSX) (Egertson *et al*, 2013) can be used to improve data selectivity by isolating and co-fragmenting at each cycle different non-contiguous precursor mass regions. (B) In the offset-windowed DIA approach, the precursor isolation window boundaries are offset by a discrete mass between consecutive cycles. For example, 25 *m/z* isolation windows get shifted by ± 12.5 *m/z* in each cycle. (C) In the "scanning quadrupole" isolation approach, termed SONAR (Moseley *et al*, 2018), the instrument continuously scans a wide precursor isolation window through the entire precursor mass range of interest, for example by scanning a mass range from 500 to 900 *m/z* using 200 × 20 *m/z* wide windows swiped with an 2 *m/z* increment.

forth between consecutive cycles in a way that the isolation windows overlap by a significant proportion (e.g. 25 *m/z* isolation windows overlapping by 12.5 *m/z*; Fig 5B). Similarly to MSX, the offset acquisition scheme results in differential precursor ion co-isolation and co-fragmentation in every second cycle. Therefore, interferences that originate from species co-selected in the differently overlapping part of the isolation window can be partitioned away and mathematically removed. This approach shows improved selectivity and lower limit of quantification compared to the standard non-overlapping approach, without extra cost related to cycle time, accumulation time, covered mass range or resolving power (https://skyline.gs.washington.edu/labkey/files/home/software/Skyline/2013-ASMS-Overlapped-DIA.pdf). However, the method still requires the use of an automated algorithm which pre-empts easy direct raw data inspection or direct raw fragment ion extraction prior to data deconvolution. Unlike MSX, this method is also possible on non-trapping instruments such as Q-TOFs.

Scanning quadrupole isolation The latest advancement of SWATH-MS was recently published under the name "SONAR" (Waters) by Moseley *et al* (2018). Instead of "stepping" through consecutive isolation windows cyclically one after the other, the quadrupole constantly "scans", for example, a 24 *m/z* wide isolation window through the entire precursor mass range of interest (Fig 5C). In other words, while with the original SWATH-MS method, a given precursor will conceptually be present in only one isolation window and absent from all the others, with the "scanning windows" one can see fragment ion signals appearing and disappearing

© 2018 The Authors

accordingly to the precursor's isotopes entrance and exit from the scanning window. This essentially adds a fourth dimension on top of retention time, mass and intensity to the SWATH-MS data. The method will require significant modifications of the existing peptide-centric analysis tools, but intuitively, this extra dimension will allow to access another level of selectivity.

Peptide-centric scoring of SWATH-MS data

Over the past 20 years, a wide variety of strategies have been developed to analyse DIA data. These strategies can be classified into two groups, depending on whether they are based on querying the data in a spectrum-centric or in a peptide-centric manner (Ting *et al*, 2015). In this tutorial, we focus on the chromatogram-based peptide-centric scoring analysis workflow for SWATH-MS data illustrated in Fig 6 and discuss specific considerations for each step. For alternative analysis strategies, we refer to recent reviews (Chapman *et al*, 2014; Bilbao *et al*, 2015).

Several software tools have been developed that support all or most steps of the workflow shown in Fig 6. These include for example OpenSWATH (Rost *et al*, 2014) (part of OpenMS, Rost *et al*, 2016b; Sturm *et al*, 2008), Skyline (MacLean *et al*, 2010), PeakView (SCIEX) and Spectronaut (Bruderer *et al*, 2015) (Biognosys). A recent comparative study (Navarro *et al*, 2016) using a ground truth data set has demonstrated remarkable agreement for the qualitative and quantitative results produced by those peptide-centric scoring tools as well as by the spectrum-centric scoring algorithm DIA-Umpire (Tsou *et al*, 2015) that was also included in the study.

Peptide query parameter (PQP) definition

The first step in a peptide-centric analysis is the definition of PQPs. While recently developed algorithms can score data without the need for empirically derived PQPs (Tsou *et al*, 2015; Ting *et al*, 2017), and thereby make the availability of prior knowledge in form of spectral libraries obsolete, empirical PQPs derived from experimental fragment ion spectra can still provide higher sensitivity, especially when working with rather large precursor isolation windows (> 5 *m/z*) during data acquisition. Depending on the sample complexity, several high-quality peak groups can be found for a given queried peptide. To quantitatively describe the frequency of this phenomenon, decoy peptides are used, which are expected to closely model the properties of the target peptides in terms of their chromatographic and mass spectrometric parameter space, but which should not be present in the sample. The decoy PQPs are generated *in silico* from the target peptide list by reversing or shuffling the amino acid sequence (Fig 6A). Decoy peptides are crucial for controlling the error rate of peptide detection and protein inference in SWATH-MS data, because they can be used to estimate the effect of random co-elution of fragment ion chromatograms in the underlying data (Rost *et al*, 2014). It is generally recommended to use the same number of fragment ions per peptide for both targets and decoys to not introduce any biases during peak scoring, because peptide queries with six co-eluting fragment ions can have different score distributions than queries with only three fragment ions.

Chromatogram extraction

The next step of a peptide-centric data analysis strategy is to use the defined PQPs to extract precursor and fragment ion chromatograms for the peptides of interest. As a peptide elutes from the column, the signal intensity of its queried fragment ions will vary synchronously following a usually Gaussian-like peptide elution profile, forming a chromatographic "peak group" that will be used to assess the peptide detection and to estimate its quantity (Fig 6B). Fragment ion chromatograms are extracted in the same manner for the target and decoy peptides. In essence, this chromatogram extraction step transforms, and thus reduces the multiplexed SWATH-MS2 spectra, to a data structure highly similar to SRM or PRM data.

To substantially improve the selectivity of peptide-centric SWATH-MS data analysis, the chromatogram extraction is usually not conducted throughout the whole chromatographic gradient length but only in a retention time window centred around the expected elution time for the queried peptide (similar to the scheduled SRM strategy). A set of shared endogenous (Parker *et al*, 2015b; Bruderer *et al*, 2016) or synthetic spike-in peptides (Escher *et al*, 2012) can be used to mediate transformation of the previously empirically determined elution times and to restrict the size of retention time extraction window as much as possible. The specific elution times of such reference peptides serve as beacons to define a normalized retention time (such as in the iRT approach; Escher *et al*, 2012) to which the elution times of query peptides can be aligned. The functionality to automatically perform a normalized indexed retention time alignment is implemented in many popular SWATH-MS data analysis tools (e.g. OpenSWATH, Skyline, PeakView and Spectronaut).

The mass tolerance or width of the ion extraction also directly impacts the selectivity of the chromatographic trace signals and thus the derived identification scores and abundance estimates. In general, an extraction width of half of the mass spectrometric peak width provides a good compromise between maximising the fraction

Figure 6. Principle of peptide-centric scoring of SWATH-MS data.
(A) Peptide-centric scoring begins with a set of peptide query parameters (PQPs), which represent retention time, precursor ion masses, fragment ion masses and fragment ion signal intensity coordinates for the target peptides (red table). PQPs are also required for decoy peptides and are generated, for example, by reversing the amino acid sequence of target peptides, while keeping the terminal amino acid (blue table). Decoy peptides are used to assess the chance that peptides which are expected to be absent in the sample may also be detected by chance. (B) Extracted ion chromatograms (XICs) are generated based on PQPs from the continuously acquired SWATH-MS2 spectra for target and decoy peptides. This results in a transformed and reduced data structure similar to data generated by targeted proteomics (SRM or PRM). (C) Fragment ion chromatograms are grouped according to their peptide association and "peak groups" with defined peak boundaries in the retention time dimension are selected. (D) For both target and decoy peak groups, a range of chromatogram- and spectrum-based scores are computed and combined to a discriminant score by a semi-supervised learning approach. The false discovery rate (FDR) of a set of detected peptides can be estimated by statistical modelling of the score distributions of target and decoy peptides. (E) For large-scale SWATH-MS analyses, error rate control should not only be performed on the peptide level, but should be extended to the protein level. Further, in large experiments including many samples, it might not be sufficient to conduct error rate control individually per run ("run-specific" context), but better on an "experiment-wide" scale. The "global" context considers only the best scoring detected peak groups, peptides or inferred proteins over all runs. (F) A multi-run alignment allows to correct or reinforce confidence in peak detection by leveraging the chromatographic time consistency and transfer of detection confidence across runs.

 © 2018 The Authors

A Peptide query parameter (PQP) definition

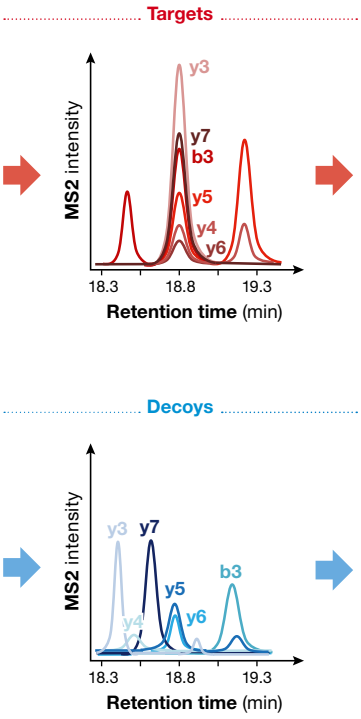
Peptide query parameters for **target peptide** `AAHTEDINACTLTTSPR`

Protein	Peptide sequence	Precursor ion *m/z*	Fragment ion *m/z*	Fragment ion type	Fragment relative intensity	Normalized retention time
Raf1	AAHTEDINACTLTTSPR	929.4416	775.4308	y7	70	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	674.3832	y6	10	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	561.2991	y5	40	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	460.2514	y4	20	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	359.2037	y3	100	45.3
Raf1	AAHTEDINACTLTTSPR	929.4416	280.1404	b3	60	45.3

Sequence reversing

Peptide query parameters for **decoy peptide** `PSTTLTCANIDETHAAR`

Decoy protein	Decoy peptide sequence	Decoy precursor ion *m/z*	Decoy fragment ion *m/z*	Decoy fragment ion type	Fragment relative intensity	Normalized retention time
Raf1	PSTTLTCANIDETHAAR	929.4416	799.3693	y7	70	45.3
Raf1	PSTTLTCANIDETHAAR	929.4416	684.3424	y6	10	45.3
Raf1	PSTTLTCANIDETHAAR	929.4416	555.2998	y5	40	45.3
Raf1	PSTTLTCANIDETHAAR	929.4416	454.2521	y4	20	45.3
Raf1	PSTTLTCANIDETHAAR	929.4416	317.1932	y3	100	45.3
Raf1	PSTTLTCANIDETHAAR	929.4416	286.1397	b3	60	45.3



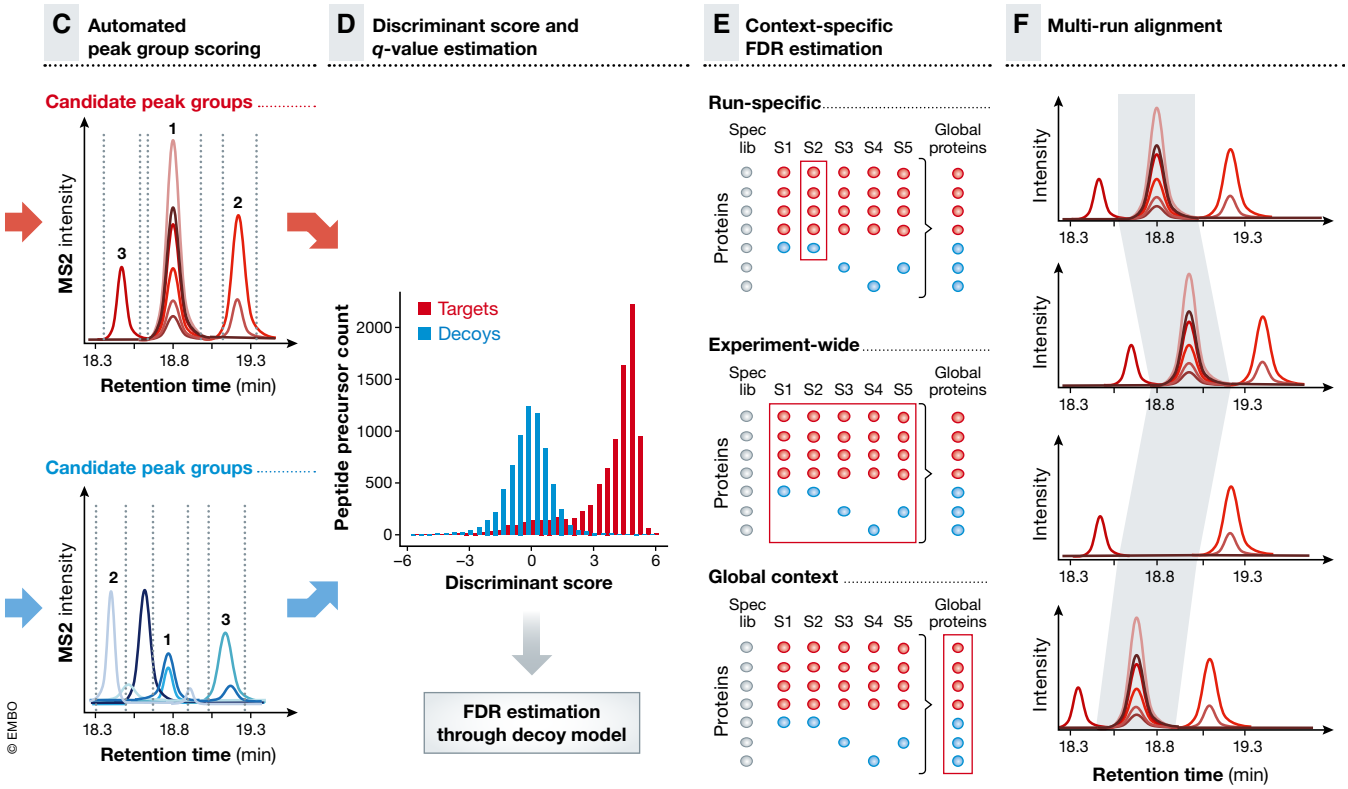


Figure 6.

of recovered ion signals while maintaining a reasonable selectivity of the extracted data. In practice, we suggest applying an extraction width of ~50 ppm or less for MS2 measurements performed with a resolution of 15,000.

Chromatographic extraction can be performed on profile (raw) or centroided MS2 data level. Centroided MS2 data are significantly smaller in size and were shown to perform similarly to profile data in terms of identification depth and quantification accuracy (Navarro *et al*, 2016). The optimal extraction window width for centroided data may be less than that of the same data extracted in profile mode, because in this case the extraction window width can be optimized with respect to the mass error distribution. However special attention should be given to the choice of centroiding algorithm used. We have observed that they can dramatically affect the data in terms of mass accuracy, signal-to-noise levels and relative fragment ion intensity (e.g. depending on whether the centroiding is done on peak height or peak area) (Toprak *et al*, 2014). Accurate centroiding is especially challenging for SWATH-MS data due to the high level of co-fragmenting peptides and the high level of overlapping fragment ion signals in the MS2 spectra. Such algorithm-specific artefacts will affect the scoring of the fragment ion peak groups as well as the statistics of peptide detection. It is thus recommended to follow the instrument vendor and algorithm developer instructions for the choice of centroiding algorithm.

Finally, if the SWATH-MS data acquisition scheme includes MS1 scans within a chromatography-compatible cycle time, precursor ion chromatograms can also be extracted from the MS1 data and scored in combination with the fragment ion peak group to enhance confidence in the peptide detection.

Automated peak group scoring

In the next step of the workflow, an algorithm defines the left and right borders of one or several potential chromatographic peak groups and scores each of these candidate peptide signals individually (Fig 6C). This is done independently for the extracted target as well as decoy chromatographic traces. During this scoring process, a variety of individual scores is computed (Reiter *et al*, 2011). Those can be categorized into five types: (i) scores related to chromatographic performance, (ii) scores correlating the measured data to the information from the spectral library, (iii) scores correlating the data to external or internal isotope-labelled standards (if available), (iv) scores taking into account the high resolution and accurate mass of the MS2 scans and (v) scores that exploit information from the precursor ion (MS1) scans. All or most of those individual scores have been implemented in one way or another into the SWATH-MS software tools OpenSWATH, Skyline, PeakView and Spectronaut, which all automatically combine them into a final discriminant score.

Discriminant score and q-value estimation

By applying the individual scores described before to extracted ion chromatograms of target and decoy peptides from SWATH-MS data, individual score distributions can be computed. Ideally, in these distributions, "true positive" target and "false positive" decoy peptides are clearly separable. However, depending on experiment-specific factors such as sample complexity, instrument performance and SWATH-MS method setup, some scores might perform better than others, while a single individual score is generally not discriminative enough to ascertain the detection and correct quantification of a given peptide. Therefore, it is beneficial to combine the individual scores into a single discriminant score in a way that allows the most sensitive recall of the peptide queries (Fig 6D). This goal can be achieved by using a semi-supervised learning algorithm (Kall *et al*, 2007) that converges iteratively towards the most optimal set of weighted individual score combinations to separate targets and decoys. The peak group with the highest discriminant score (rank 1, Fig 6C) is then usually considered as the most likely detected peak group for a set of queried peptides, which may eventually be revised upon across-run alignment.

Several requirements need to be fulfilled in order to allow successful learning and proper statistical modelling of SWATH-MS data. First, a sufficient number of high scoring true positive peptides are necessary for the semi-supervised algorithm to learn the scoring characteristics of true detection events and to achieve a sensible combination of individual weight-feature scores. Second, the decoy peptides must appropriately represent peptides that are not detectable in the sample.

In general, we highly recommend to inspect target and decoy score distribution plots that most SWATH-MS analysis tools report to verify that (i) assumptions made with regard to the decoy distributions (e.g. normality) are not violated, (ii) the target scores appear approximately as a bimodal distribution when analysing complex samples with medium to large spectral libraries and (iii) the decoy distribution matches the true-negative part of the bimodal target distribution both in terms of apex, shape and width (Fig 6D). In scenarios, where the spectral library overlaps to a very high degree (> 90%) with the detected precursor ions of the sample, the true-negative target distribution might be very small, thus rendering criterion (iii) difficult to judge (Reiter *et al*, 2011). If one or more of those conditions are not fulfilled, it is likely that either the machine learning or the q-value estimation step has failed or are biased, which could result in misleading statistics and inappropriate error rate control.

Context-specific error rate estimation

The above described strategies are commonly applied to individual runs separately. However, under certain circumstances, this might introduce biases, especially when large numbers of peptides are queried across data sets consisting of a large number of heterogeneous samples, or when spectral libraries are used that were assembled from different sample types (e.g. obtained from public repositories) or contain a large number of true-negative peptides and proteins for the tested samples. We therefore suggest to adapt the analysis strategy according to requirements of the specific data set as described recently by Rosenberger *et al* (2017a). Briefly, to ensure comparable scoring and statistics throughout the complete data set, the semi-supervised machine learning step should be conducted in an experiment-wide fashion on all the candidate peak groups for all runs at once (Fig 6E). For practical reasons (file size and memory restrictions), this step can also be performed on a representative selection of peptide queries randomly subsampled through each sample of the data set. Then, the same score weights are applied to all the candidate peak groups throughout all the runs of the data set (Rosenberger *et al*, 2017a).

While a 1% peptide FDR threshold on a per run basis (referred to as "run-specific" context) might be sufficient for small sample sizes

 © 2018 The Authors

analysed with sample-specific libraries, accumulation of false positives will occur when conducting large numbers of queries across large sample sets. Similar as for DDA database search strategies, more stringent score cut-offs become necessary on large data sets to account for multiple hypothesis testing. We therefore recommend to also use target and decoy score distributions to conduct error rate control at the protein level in addition to the peptide level (Rosenberger *et al*, 2017a). Further, we suggest assessing the FDR in an "experiment-wide" context where q-values are estimated over all samples using a single score distribution, as well as in a "global" context, where only the best scoring instance of a given peptide or inferred protein from all samples in the experiment is used. We recommend that the list of peptide queries and inferred proteins identified in the "global" context at 1% FDR can be used to filter the resulting data from the "experiment-wide" context to prevent accumulation of false positives in experiments with many samples (Fig 6E; Rosenberger *et al*, 2017a). The above described context-specific error estimation has been implemented in PyProphet (Rosenberger *et al*, 2017a), a component of the OpenSWATH workflow, and Spectronaut (Bruderer *et al*, 2017).

Across-run alignment and transfer of identification confidence

A main assumption of peptide-centric scoring is that only the best scoring peak group per peptide represents the true peptide signal. This may however not always be the case, for example when the peptide abundance drops below the limit of detection in one or several runs of a sample cohort and a candidate peak group that does not originate from the target peptide achieves a better score. Frequently, this wrong peak group is detected at a different retention time than the true peptide, resulting in a false value in the final peptide quantification matrix. By re-emphasizing the consistency of the retention time information across samples and aligning the peak boundaries of the peak group from the run where the peptide was most confidently identified, it is possible to detect and amend those false peak group ranking mistakes (Fig 6F). This process sometimes re-ranks the peak groups or may rescue some peak groups at the expected retention time that would otherwise not have passed the 1% peptide FDR threshold in those specific runs. An algorithm for multi-run retention time alignment is TRIC (TRansfer of Identification Confidence), a component of the OpenSWATH workflow (Rost *et al*, 2016a).

In general, multi-run alignment is similar to what has already been developed and applied successfully for the analysis of DDA data (Prakash *et al*, 2006; Mueller *et al*, 2007; Cox *et al*, 2014). However, while in DDA peptide identification and quantification are often inferred across runs based solely on retention time and precursor *m/z*, in SWATH-MS the full information of all fragment ions can be used as part of the alignment strategy and to control the error rate.

Peptide and protein quantification

Just like any other bottom-up LC-MS/MS proteomic methods, SWATH-MS provides quantification data on peptide level and peptide quantities are typically computed by summing or averaging the integrated peak area of several fragment ions. It is not necessary to use all of the queried fragment ions per peptide to perform accurate quantification. Rather only high-quality and interference-free fragment traces should be selected and for such a task automated algorithms have been developed (Keller *et al*, 2015; Teleman *et al*, 2015). Additionally, MS1-based quantification on the precursor extraction ion chromatograms may provide orthogonal quantification information (Rardin *et al*, 2015). In many instances, however, the lower limit of quantification (LLOQ) appears to be more strongly impaired at the precursor than at the fragment ion level (Gillet *et al*, 2012; Egertson *et al*, 2013; Collins *et al*, 2017). This is due to the lower selectivity and the higher dynamic range of detection required in full MS1 scans compared to that of windowed-isolation MS2 scans.

To infer the abundance of a protein from a SWATH-MS measurement, the measured intensities of one or many peptides need to be aggregated into a final protein intensity value per sample. For this purpose, various strategies have been developed, for example (i) summing or averaging the n-most intense peptides per protein (Silva *et al*, 2006; Ludwig *et al*, 2012), (ii) summing up all peptide intensities per sample, regardless of whether they occur across samples or not (iBAQ; Schwanhausser *et al*, 2011), or (iii) considering only those peptides per protein that occur in the two samples to be compared (LFQ; Cox *et al*, 2014). Optionally, those aggregated protein intensity values can be further normalized by a factor representative of the detectability of that protein, for example protein length (Zybailov *et al*, 2006), number of theoretical peptides generated (Schwanhausser *et al*, 2011) or a computed peptide detection probability score (Lu *et al*, 2007). Missing values can be either recovered using the re-quantification values provided from the across-run re-alignment analysis (Rost *et al*, 2016a) or using statistical imputation as provided by other tools (Karpievitch *et al*, 2012). Compared to SRM or PRM, which commonly only use a few measured targeted peptides for protein inference, and to DDA, which may not identify and quantify the same peptides per protein through large sample sets, SWATH-MS peptide-centric analysis intrinsically offers a more consistent and complete choice in peptides for protein quantification.

Several statistical tools have been developed to assess differential peptide or protein abundance between experimental groups and to compute the relative fold change in peptide or protein quantification starting from fragment ion level data (Chang *et al*, 2012). For most consistent relative protein quantification comparisons, we recommend to systematically use the same, and throughout the complete SWATH-MS data set the most robust, fragments and peptides per protein. For the selection of these fragments and peptides, information such as conservation of the relative fragment ion signal intensities per peptide and the relative peptide intensities per protein can be leveraged across the whole data set, as MSstats (Choi *et al*, 2014) or mapDIA (Teo *et al*, 2015) do for fold-change statistics computation.

Post-translational modifications

SWATH-MS peptide-centric scoring is directly compatible with measuring peptides carrying post-translational modifications (PTMs). However, analysis of modified peptides by SWATH-MS presents additional challenges that stem from the often high similarity of MS2 spectra arising from related peptide species, including the non-modified peptide, modified peptides with the same backbone carrying the same modification at different sites or peptides with the same backbone carrying (near) isobaric modifications. Such peptides may be co-isolated in the same SWATH-MS window and produce multiple high scoring peak groups.

As in DDA analysis, specific strategies have been developed to confidently identify or detect modified peptides in DIA and

© 2018 The Authors

SWATH-MS data sets. The most straightforward way is to directly use PQPs of the modified peptide versions, for example from spectral libraries that were acquired from DDA analyses of PTM-enriched samples or with open modification search strategies (Na & Paek, 2015). An alternative strategy relies on using non-modified PQPs to query for modified peptides and focuses on additional high scoring peak groups. Those additional peak groups may be detected in the same isolation window as the non-modified peptides or in other isolation windows. Such strategies have been implemented in MSPLIT-DIA (Wang *et al*, 2015) and SWATHProphet^{PTM} (Keller *et al*, 2016) and were shown to work most successfully when the peptide fragmentation shows enough similarity up to the site of the modification (Toprak *et al*, 2014).

A recently introduced algorithm termed "Inference of PeptidoForms" (IPF) (Rosenberger *et al*, 2017b) extends the standard OpenSWATH workflow to modified peptides. IPF uses a two-stage workflow in which the first pass resembles a standard OpenSWATH analysis whereby typically, about six fragment ions from DDA- or DIA-derived spectral libraries are extracted and scored to detect high-quality peak groups. These transitions need to be specific for the targeted peptide sequence and types and numbers of modifications; however, they do not need to be specific for positional isomers. In a second step, IPF uses XICs extracted both for precursor signals in MS1 and theoretically predicted fragment ions that can differentiate positional isomers. This information is then integrated using a Bayesian hierarchical model leading to a single peptidoform confidence score for each detected peak group. The site localization is thus independently reassessed, even if the underlying peptidoform is not present in the spectral library.

In summary, PTM-SWATH analyses are applicable to total cellular lysates as well as to samples enriched for specific types of modifications, such as phospho-enriched samples. In recent years, several different PTM-types have been successfully studied with SWATH-MS (Krautkramer *et al*, 2015; Sidoli *et al*, 2015; Lawrence *et al*, 2016; Rosenberger *et al*, 2017b). Particularly in experiments involving many experimental conditions or replicates, similar improvements in terms of consistent detection and quantification can be expected for modified and unmodified peptides.

Outlook

Since the first implementations of DIA in the early 2000s (Masselon *et al*, 2000; Purvine *et al*, 2003; Venable *et al*, 2004), numerous improvements in terms of data acquisition speed, mass accuracy and resolution have taken place on the instrument side. Also, since the first description of a peptide-centric data analysis workflow (Gillet *et al*, 2012), substantial progress has been achieved in terms of spectral library generation (Schubert *et al*, 2015a), automated data analysis pipelines (Rost *et al*, 2014, 2016a; Keller *et al*, 2015) and statistical control in peptide and protein error rates (Rosenberger *et al*, 2017a). The excellent reproducibility and accuracy of SWATH-MS data acquisition have been proven in a large study comparing data acquired in different laboratories worldwide (Collins *et al*, 2017). Also, the robustness of peptide-centric data analysis tools has been demonstrated recently by comparing five major software tools currently used for peptide-centric scoring of SWATH-MS data (Navarro *et al*, 2016). The development of SWATH-MS laid the foundation for novel large-scale biological studies, such as the study of mitochondrial links to liver metabolism in a mouse reference population (Williams *et al*, 2016), the study of phosphoproteomic and chromatin signature in response to drug treatments in cancer cell lines (preprint: Litichevskiy *et al*, 2018) or the contribution of heritability and environment to different traits in pairs of monozygotic and dizygotic twins (Liu *et al*, 2015). Throughout this tutorial, we have documented the advancements and referenced established workflows, software tools and analysis pipelines that will hopefully support and facilitate SWATH-MS experiments by a large user community in the years to come.

Although most published SWATH-MS studies to date used spectral libraries generated by DDA, novel approaches that do not rely on libraries were developed. Spectrum-centric scoring approaches such as DIA-Umpire (Tsou *et al*, 2015) or Group-DIA (Li *et al*, 2015) generate pseudo-MS2 spectra directly from DIA data and subject them to conventional database search algorithms. Peptide-centric library-free analysis tools such as FT-ARM (Weisbrod *et al*, 2012) or PECAN (Ting *et al*, 2017) use theoretical fragment ion predictions to query and score the multiplexed MS2 spectra of SWATH-MS data. Interestingly, tools that rely on prior knowledge in the form of spectral libraries seem to deal better with lower selectivity data than library-free tools (Navarro *et al*, 2016; Ting *et al*, 2017). This finding suggests that, as measurement selectivity improves with better instruments (smaller precursor isolation windows, higher peak capacity chromatography, scanning quadrupoles, ion mobility separations, etc.), prior knowledge for data analysis might become less important. On the other hand, prior knowledge in the form of organism-wide deep proteome data sets from endogenous samples or synthetic proteomics is likely to increase substantially during the same time period. This might spur the development of novel analysis methods that can better leverage such libraries.

In the future, we expect that mass spectrometers will continue to improve in sensitivity and scanning speed. Hence, it is foreseeable that the time necessary to acquire high-quality precursor and fragment ion spectra will decrease, allowing for higher number of narrower isolation windows and enabling to approach the 1–3 *m/z* isolation width commonly used during DDA acquisition and PRM. We speculate that increasing instrument sensitivity and speed will reach a threshold where ultimately the benefits of all currently employed data acquisition methods can be achieved with a single "super" method. The question then will no longer be whether one should do a DDA, DIA or targeted experiment, but whether to use a peptide-centric or spectrum-centric data analysis strategy. It seems unlikely that a purely spectrum-centric approach, where each consecutive MS2 scan is subjected to a database search algorithm independently, would be most effective, because the rich information in the chromatographic dimension and prior information about the fragmentation pattern would not be leveraged. As such, it seems more likely that either peptide-centric approaches as described in detail in this tutorial or hybrid methods between spectrum- and peptide-centric analyses best exploit the highly comprehensive and selective data sets in the future.

Other future-oriented features of DIA in general, and SWATH-MS in particular, are the simplicity of the underlying data acquisition method and the suitability for high-throughput proteomics. Already today ~50% of the MS-detectable proteome can be reproducibly measured by SWATH-MS at a relatively fast time scale (< 1 h per

 © 2018 The Authors

sample). This progress will hopefully allow to further democratize access to proteomics data in the future and to perform novel types of cross study comparisons, which are goals that the proteomics community had difficulties achieving in the past.

Expanded View for this article is available online.

Acknowledgements
We would like to acknowledge Isabell Bludau for providing Fig 6E. We also thank Olga Schubert and Hannes Röst for proofreading the manuscript. We want to express our gratitude to the many course participants and speakers of the "targeted proteomics" and "DIA/SWATH" courses taking place at ETH Zürich over the last 5 years, who inspired and fuelled this tutorial with lots of discussions and their continuous input and interest. C.L. would like to especially thank Kim, Lilli, Linus and Hildegard Ludwig, for their continuous support during the long writing process of this manuscript. B.C.C. was supported by a Swiss National Science Foundation Ambizione Grant (PZ00P3_161435). R.A. was supported by ERC Proteomics v3.0 (AdG-233226) and Proteomics 4D (AdG-670821), the PhosphoNetX Project of SystemsX.ch and the Swiss National Science Foundation (SNSF) grant number 31003A_166435. L.G. was supported by ERC Proteomics 4D (AdG-670821) and H2020 – PrECISE (668858).

Conflict of interest
The authors declare that they have no conflict of interest. RA is a shareholder in the company Biognosys which operates in the field of research covered by this article.

References

Aebersold R, Mann M (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* 537: 347 – 355

Ahrne E, Glatter T, Vigano C, Schubert C, Nigg EA, Schmidt A (2016) Evaluation and improvement of quantification accuracy in isobaric mass tag-based protein quantification experiments. *J Proteome Res* 15: 2537 – 2547

Bereman MS, Beri J, Sharma V, Nathe C, Eckels J, MacLean B, MacCoss MJ (2016) An automated pipeline to monitor system performance in liquid chromatography-tandem mass spectrometry proteomic experiments. *J Proteome Res* 15: 4763 – 4769

Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, Muller M, Lisacek F (2015) Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15: 964 – 980

Blainey P, Krzywinski M, Altman N (2014) Points of significance: replication. *Nat Methods* 11: 879 – 880

Bruderer R, Bernhardt OM, Gandhi T, Miladinovic SM, Cheng LY, Messner S, Ehrenberger T, Zanotelli V, Butscheid Y, Escher C, Vitek O, Rinner O, Reiter L (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* 14: 1400 – 1410

Bruderer R, Bernhardt O, Gandhi T, Reiter L (2016) High precision iRT retention time prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* 16: 2246 – 2256

Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, Schmidt M, Gomez-Varela D, Reiter L (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol Cell Proteomics* 16: 2296 – 2309

Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, Deutsch EW, Grant RP, Hoofnagle AN, Huttenhain R, Koomen JM, Liebler DC, Liu T, MacLean B, Mani DR, Mansfield E, Neubert H, Paulovich AG, Reiter L, Vitek O, Aebersold R *et al* (2014) Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol Cell Proteomics* 13: 907 – 917

Chang CY, Picotti P, Huttenhain R, Heinzelmann-Schwarz V, Jovanovic M, Aebersold R, Vitek O (2012) Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol Cell Proteomics* 11: M111.014662

Chapman JD, Goodlett DR, Masselon CD (2014) Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* 33: 452 – 470

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP (2016) Defining the consequences of genetic variation on a proteome–wide scale. *Nature* 534: 500 – 505

Chiva C, Olivella R, Borras E, Espadas G, Pastor O, Sole A, Sabido E (2018) QCloud: a cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS ONE* 13: e0189209

Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 30: 2524 – 2526

Collins BC, Gillet LC, Rosenberger G, Rost HL, Vichalkovski A, Gstaiger M, Aebersold R (2013) Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat Methods* 10: 1246 – 1253

Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, Chan DW, Gibson BW, Gingras AC, Held JM, Hirayama-Kurogi M, Hou G, Krisp C, Larsen B, Lin L, Liu S, Molloy MP, Moritz RL, Ohtsuki S, Schlapbach R *et al* (2017) Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun* 8: 291

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13: 2513 – 2526

Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 3: 1234 – 1242

Deutsch EW, Lam H, Aebersold R (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 9: 429 – 434

Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R (2010) A guided tour of the trans-proteomic pipeline. *Proteomics* 10: 1150 – 1159

Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V, Wu CC, MacCoss MJ (2013) Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* 10: 744 – 746

Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ (2015) Multiplexed peptide analysis using data-independent acquisition and skyline. *Nat Protoc* 10: 887 – 903

Escher C, Reiter L, MacLean B, Ossola R, Herzog F, Chilton J, MacCoss MJ, Rinner O (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 12: 1111 – 1121

© 2018 The Authors

Frejno M, Zenezini Chiozzi R, Wilhelm M, Koch H, Zheng R, Klaeger S, Ruprecht B, Meng C, Kramer K, Jarzab A, Heinzlmeir S, Johnstone E, Domingo E, Kerr D, Jesinghaus M, Slotta-Huspenina J, Weichert W, Knapp S, Feller SM, Kuster B (2017) Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol Syst Biol* 13: 951

Geiger T, Cox J, Mann M (2010) Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics* 9: 2252 – 2261

Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M (2016) Plasma proteome profiling to assess human health and disease. *Cell Syst* 2: 185 – 195

Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 11: O111.016717

de Graaf EL, Altelaar AF, van Breukelen B, Mohammed S, Heck AJ (2011) Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra. *J Proteome Res* 10: 4334 – 4341

Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 4: 923 – 925

Karlsson C, Malmstrom L, Aebersold R, Malmstrom J (2012) Proteome-wide selected reaction monitoring assays for the human pathogen Streptococcus pyogenes. *Nat Commun* 3: 1301

Karpievitch YV, Dabney AR, Smith RD (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 13(Suppl 16): S5

Keam SP, Gulati T, Gamell C, Caramia F, Huang C, Schittenhelm RB, Kleifeld O, Neeson PJ, Haupt Y, Williams SG (2018) Exploring the oncoproteomic response of human prostate cancer to therapeutic radiation using data-independent acquisition (DIA) mass spectrometry. *Prostate* 78: 563 – 575

Keller A, Bader SL, Shteynberg D, Hood L, Moritz RL (2015) Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. *Mol Cell Proteomics* 14: 1411 – 1418

Keller A, Bader SL, Kusebauch U, Shteynberg D, Hood L, Moritz RL (2016) Opening a SWATH window on posttranslational modifications: automated pursuit of modified peptides. *Mol Cell Proteomics* 15: 1151 – 1163

Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV (2018) Performance evaluation of the Q exactive HF-X for shotgun proteomics. *J Proteome Res* 17: 727 – 738

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 14: 513 – 520

Krautkramer KA, Reiter L, Denu JM, Dowell JA (2015) Quantification of SAHA-dependent changes in histone modifications using data-independent acquisition mass spectrometry. *J Proteome Res* 14: 3252 – 3262

Krzywinski M, Altman N (2013) Points of significance: power and sample size. *Nat Methods* 10: 1139 – 1140

Krzywinski M, Altman N (2014a) Points of significance: analysis of variance and blocking. *Nat Methods* 11: 699 – 700

Krzywinski M, Altman N (2014b) Points of view: designing comparative experiments. *Nat Methods* 11: 597 – 598

Kulkarni S, Koller A, Mani KM, Wen R, Alfieri A, Saha S, Wang J, Patel P, Bandeira N, Guha C, Chen EI (2016) Identifying urinary and serum exosome biomarkers for radiation exposure using a data dependent acquisition and SWATH-MS combined workflow. *Int J Radiat Oncol Biol Phys* 96: 566 – 577

Kusebauch U, Campbell DS, Deutsch EW, Chu CS, Spicer DA, Brusniak MY, Slagel J, Sun Z, Stevens J, Grimes B, Shteynberg D, Hoopmann MR, Blattmann P, Ratushny AV, Rinner O, Picotti P, Carapito C, Huang CY, Kapousouz M, Lam H *et al* (2016) Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell* 166: 766 – 778

Kuster B, Schirle M, Mallick P, Aebersold R (2005) Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* 6: 577 – 583

Lambert JP, Ivosev G, Couzens AL, Larsen B, Taipale M, Lin ZY, Zhong Q, Lindquist S, Vidal M, Aebersold R, Pawson T, Bonner R, Tate S, Gingras AC (2013) Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat Methods* 10: 1239 – 1245

Lange V, Picotti P, Domon B, Aebersold R (2008) Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* 4: 222

Lawrence RT, Perez EM, Hernandez D, Miller CP, Haas KM, Irie HY, Lee SI, Blau CA, Villen J (2015) The proteomic landscape of triple-negative breast cancer. *Cell Rep* 11: 630 – 644

Lawrence RT, Searle BC, Llovet A, Villen J (2016) Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat Methods* 13: 431 – 434

Li Y, Zhong CQ, Xu X, Cai S, Wu X, Zhang Y, Chen J, Shi J, Lin S, Han J (2015) Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat Methods* 12: 1105 – 1106

Litichevskiy L, Peckner R, Abelin JG, Asiedu JK, Creech AL, Davis JF, Davison D, Dunning CM, Egertson JD, Egri S, Gould J, Ko T, Johnson SA, Lahr DL, Lam D, Liu Z, Lyons NJ, Lu X, MacLean BX, Mungenast AE *et al* (2018) A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. *Cell Syst* 6: 424 – 443.e7

Liu Y, Huttenhain R, Surinova S, Gillet LC, Mouritsen J, Brunner R, Navarro P, Aebersold R (2013) Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics* 13: 1247 – 1256

Liu Y, Chen J, Sethi A, Li QK, Chen L, Collins B, Gillet LC, Wollscheid B, Zhang H, Aebersold R (2014) Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. *Mol Cell Proteomics* 13: 1753 – 1768

Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, Vitek O, Mouritsen J, Lachance G, Spector TD, Dermitzakis ET, Aebersold R (2015) Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 11: 786

Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25: 117 – 124

Ludwig C, Claassen M, Schmidt A, Aebersold R (2012) Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol Cell Proteomics* 11: M111.013987

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26: 966 – 968

Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 25: 125 – 131

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R (2005) PRIDE: the proteomics identifications database. *Proteomics* 5: 3537 – 3545

 © 2018 The Authors

Masselon C, Anderson GA, Harkewicz R, Bruce JE, Pasa-Tolic L, Smith RD (2000) Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures. *Anal Chem* 72: 1918 – 1924

Matsumoto M, Matsuzaki F, Oshikawa K, Goshima N, Mori M, Kawamura Y, Ogawa K, Fukuda E, Nakatsumi H, Natsume T, Fukui K, Horimoto K, Nagashima T, Funayama R, Nakayama K, Nakayama KI (2017) A large-scale targeted proteomics assay resource based on an *in vitro* human proteome. *Nat Methods* 14: 251 – 258

Matthews DE, Hayes JM (1976) Systematic errors in gas chromatography-mass spectrometry isotope ratio measurements. *Anal Chem* 48: 1375 – 1382

McAlister GC, Huttlin EL, Haas W, Ting L, Jedrychowski MP, Rogers JC, Kuhn K, Pike I, Grothe RA, Blethrow JD, Gygi SP (2012) Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* 84: 7469 – 7478

McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP (2014) MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* 86: 7150 – 7158

Moseley MA, Hughes CJ, Juvvadi PR, Soderblom EJ, Lennon S, Perkins SR, Thompson JW, Steinbach WJ, Geromanos SJ, Wildgoose J, Langridge JI, Richardson K, Vissers JPC (2018) Scanning quadrupole data-independent acquisition, part A: qualitative and quantitative characterization. *J Proteome Res* 17: 770 – 779

Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Muller M (2007) SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7: 3470 – 3480

Muller DB, Schubert OT, Rost H, Aebersold R, Vorholt JA (2016) Systems-level proteomics of two ubiquitous leaf commensals reveals complementary adaptive traits for phyllosphere colonization. *Mol Cell Proteomics* 15: 3256 – 3269

Muntel J, Xuan Y, Berger ST, Reiter L, Bachur R, Kentsis A, Steen H (2015) Advancing urinary protein biomarker discovery by data-independent acquisition on a quadrupole-orbitrap mass spectrometer. *J Proteome Res* 14: 4752 – 4762

Na S, Paek E (2015) Software eyes for protein post-translational modifications. *Mass Spectrom Rev* 34: 133 – 147

Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Rost HL, Tate SA, Tsou CC, Reiter L, Distler U, Rosenberger G, Perez-Riverol Y, Nesvizhskii AI, Aebersold R, Tenzer S (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* 34: 1130 – 1136

Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *J Am Soc Mass Spectrom* 16: 1239 – 1249

O'Brien JJ, O'Connell JD, Paulo JA, Thakurta S, Rose CM, Weekes MP, Huttlin EL, Gygi SP (2018) Compositional proteomics: effects of spatial constraints on protein quantification utilizing isobaric tags. *J Proteome Res* 17: 590 – 599

Okada H, Ebhardt HA, Vonesch SC, Aebersold R, Hafen E (2016) Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in *Drosophila melanogaster*. *Nat Commun* 7: 12649

Ortea I, Rodriguez-Ariza A, Chicano-Galvez E, Arenas Vacas MS, Jurado Gamez B (2016) Discovery of potential protein biomarkers of lung adenocarcinoma in bronchoalveolar lavage fluid by SWATH MS data-independent acquisition and targeted data extraction. *J Proteomics* 138: 106 – 114

Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR (2009) Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem* 81: 6481 – 6488

Panse C, Trachsel C, Grossmann J, Schlapbach R (2015) specL–an R/Bioconductor package to prepare peptide spectrum matches for use in targeted proteomics. *Bioinformatics* 31: 2228 – 2231

Parker BL, Yang G, Humphrey SJ, Chaudhuri R, Ma X, Peterman S, James DE (2015a) Targeted phosphoproteomics of insulin signaling using data-independent acquisition mass spectrometry. *Sci Signal* 8: rs6

Parker SJ, Rost H, Rosenberger G, Collins BC, Malmstrom L, Amodei D, Venkatraman V, Raedschelders K, Van Eyk JE, Aebersold R (2015b) Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry. *Mol Cell Proteomics* 14: 2800 – 2813

Picotti P, Lam H, Campbell D, Deutsch EW, Mirzaei H, Ranish J, Domon B, Aebersold R (2008) A database of mass spectrometric assays for the yeast proteome. *Nat Methods* 5: 913 – 914

Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, Wenschuh H, Aebersold R (2010) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat Methods* 7: 43 – 46

Picotti P, Clement-Ziza M, Lam H, Campbell DS, Schmidt A, Deutsch EW, Rost H, Sun Z, Rinner O, Reiter L, Shen Q, Michaelson JJ, Frei A, Alberti S, Kusebauch U, Wollscheid B, Moritz RL, Beyer A, Aebersold R (2013) A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* 494: 266 – 270

Prakash A, Mallick P, Whiteaker J, Zhang H, Paulovich A, Flory M, Lee H, Aebersold R, Schwikowski B (2006) Signal maps for mass spectrometry-based comparative proteomics. *Mol Cell Proteomics* 5: 423 – 432

Purvine S, Eppel JT, Yi EC, Goodlett DR (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* 3: 847 – 850

Rardin MJ, Schilling B, Cheng LY, MacLean BX, Sorensen DJ, Sahu AK, MacCoss MJ, Vitek O, Gibson BW (2015) MS1 peptide ion intensity chromatograms in MS2 (SWATH) data independent acquisitions. Improving post acquisition analysis of proteomic experiments. *Mol Cell Proteomics* 14: 2405 – 2419

Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, Aebersold R (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 8: 2405 – 2417

Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, Hengartner MO, Aebersold R (2011) mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat Methods* 8: 430 – 435

Rosenberger G, Koh CC, Guo T, Rost HL, Kouvonen P, Collins BC, Heusel M, Liu Y, Caron E, Vichalkovski A, Faini M, Schubert OT, Faridi P, Ebhardt HA, Matondo M, Lam H, Bader SL, Campbell DS, Deutsch EW, Moritz RL *et al* (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* 1: 140031

Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, Liu Y, MacCoss MJ, MacLean BX, Nesvizhskii AI, Pedrioli PGA, Reiter L, Rost HL, Tate S, Ting YS, Collins BC, Aebersold R (2017a) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Methods* 14: 921 – 927

Rosenberger G, Liu Y, Rost HL, Ludwig C, Buil A, Bensimon A, Soste M, Spector TD, Dermitzakis ET, Collins BC, Malmstrom L, Aebersold R (2017b) Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat Biotechnol* 35: 781 – 788

© 2018 The Authors

Rost H, Malmstrom L, Aebersold R (2012) A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol Cell Proteomics* 11: 540 – 549

Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L, Aebersold R (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 32: 219 – 223

Rost HL, Liu Y, D'Agostino G, Zanella M, Navarro P, Rosenberger G, Collins BC, Gillet L, Testa G, Malmstrom L, Aebersold R (2016a) TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat Methods* 13: 777 – 783

Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E, Liang X, Nahnsen S, Nilse L, Pfeuffer J, Rosenberger G, Rurik M, Schmitt U, Veit J, Walzer M, Wojnar D *et al* (2016b) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* 13: 741 – 748

Roumeliotis TI, Williams SP, Gonçalves E, Alsinet C, Del Castillo Velasco-Herrera M, Aben N, Ghavidel FZ, Michaut M, Schubert M, Price S, Wright JC, Yu L, Yang M, Dienstmann R, Guinney J, Beltrao P, Brazma A, Pardo M, Stegle O, Adams DJ *et al* (2017) Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell Rep* 20: 2201 – 2214

Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham AJ, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM *et al* (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics* 9: 225 – 241

Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, Becher I, Pachl F, Kuster B, Bantscheff M (2013) Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J Proteome Res* 12: 3586 – 3598

Schmidlin T, Garrigues L, Lane CS, Mulder TC, van Doorn S, Post H, de Graaf EL, Lemeer S, Heck AJ, Altelaar AF (2016) Assessment of SRM, MRM and DIA for the targeted analysis of phosphorylation dynamics in non-small cell lung cancer. *Proteomics* 16: 2193 – 2205

Schubert OT, Mouritsen J, Ludwig C, Rost HL, Rosenberger G, Arthur PK, Claassen M, Campbell DS, Sun Z, Farrah T, Gengenbacher M, Maiolica A, Kaufmann SH, Moritz RL, Aebersold R (2013) The Mtb proteome library: a resource of assays to quantify the complete proteome of Mycobacterium tuberculosis. *Cell Host Microbe* 13: 602 – 612

Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, Lam H, Amodei D, Mallick P, MacLean B, Aebersold R (2015a) Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat Protoc* 10: 426 – 441

Schubert OT, Ludwig C, Kogadeeva M, Zimmermann M, Rosenberger G, Gengenbacher M, Gillet LC, Collins BC, Rost HL, Kaufmann SH, Sauer U, Aebersold R (2015b) Absolute proteome composition and dynamics during dormancy and resuscitation of *Mycobacterium tuberculosis*. *Cell Host Microbe* 18: 96 – 108

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* 473: 337 – 342

Searle BC, Egertson JD, Bollinger JG, Stergachis AB, MacCoss MJ (2015) Using data independent acquisition (DIA) to model high-responding peptides for targeted proteomics experiments. *Mol Cell Proteomics* 14: 2331 – 2340

Selevsek N, Chang CY, Gillet LC, Navarro P, Bernhardt OM, Reiter L, Cheng LY, Vitek O, Aebersold R (2015) Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-mass spectrometry. *Mol Cell Proteomics* 14: 739 – 749

Sharma V, Eckels J, Taylor GK, Shulman NJ, Stergachis AB, Joyner SA, Yan P, Whiteaker JR, Halusa GN, Schilling B, Gibson BW, Colangelo CM, Paulovich AG, Carr SA, Jaffe JD, MacCoss MJ, MacLean B (2014) Panorama: a targeted proteomics knowledge base. *J Proteome Res* 13: 4205 – 4210

Sidoli S, Lin S, Xiong L, Bhanu NV, Karch KR, Johansen E, Hunter C, Mollah S, Garcia BA (2015) Sequential window acquisition of all theoretical mass spectra (SWATH) analysis for characterization and quantification of histone post-translational modifications. *Mol Cell Proteomics* 14: 2420 – 2428

Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* 5: 144 – 156

Sonnett M, Yeung E, Wuhr M (2018) Accurate, sensitive, and precise multiplexed proteomics using the complement reporter ion cluster. *Anal Chem* 90: 5032 – 5039

Stergachis AB, MacLean B, Lee K, Stamatoyannopoulos JA, MacCoss MJ (2011) Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat Methods* 8: 1041 – 1043

Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* 9: 163

Tan SLW, Chadha S, Liu Y, Gabasova E, Perera D, Ahmed K, Constantinou S, Renaudin X, Lee M, Aebersold R, Venkitaraman AR (2017) A class of environmental and endogenous toxins induces BRCA2 haploinsufficiency and genome instability. *Cell* 169: 1105 – 1118.e15

Teleman J, Rost HL, Rosenberger G, Schmitt U, Malmstrom L, Malmstrom J, Levander F (2015) DIANA–algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* 31: 555 – 562

Teleman J, Hauri S, Malmstrom J (2017) Improvements in mass spectrometry assay library generation for targeted proteomics. *J Proteome Res* 16: 2384 – 2392

Teo G, Kim S, Tsou CC, Collins B, Gingras AC, Nesvizhskii AI, Choi H (2015) mapDIA: preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J Proteomics* 129: 108 – 120

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 75: 1895 – 1904

Ting L, Rad R, Gygi SP, Haas W (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* 8: 937 – 940

Ting YS, Egertson JD, Payne SH, Kim S, MacLean B, Kall L, Aebersold R, Smith RD, Noble WS, MacCoss MJ (2015) Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol Cell Proteomics* 14: 2301 – 2307

Ting YS, Egertson JD, Bollinger JG, Searle BC, Payne SH, Noble WS, MacCoss MJ (2017) PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods* 14: 903 – 908

Toprak UH, Gillet LC, Maiolica A, Navarro P, Leitner A, Aebersold R (2014) Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol Cell Proteomics* 13: 2056 – 2071

Tsou CC, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras AC, Nesvizhskii AI (2015) DIA-Umpire: comprehensive computational

 © 2018 The Authors

framework for data-independent acquisition proteomics. *Nat Methods* 12: 258 – 264

Unwin RD, Pierce A, Watson RB, Sternberg DW, Whetton AD (2005) Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. *Mol Cell Proteomics* 4: 924 – 935

Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1: 39 – 45

Vowinckel J, Zelezniak A, Bruderer R, Mulleder M, Reiter L, Ralser M (2018) Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. *Sci Rep* 8: 4346

Wang X, Chambers MC, Vega-Montoto LJ, Bunk DM, Stein SE, Tabb DL (2014) QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal Chem* 86: 2497 – 2509

Wang J, Tucholska M, Knight JD, Lambert JP, Tate S, Larsen B, Gingras AC, Bandeira N (2015) MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat Methods* 12: 1106 – 1108

Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE (2012) Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J Proteome Res* 11: 1621 – 1632

Whiteaker JR, Halusa GN, Hoofnagle AN, Sharma V, MacLean B, Yan P, Wrobel JA, Kennedy J, Mani DR, Zimmerman LJ, Meyer MR, Mesri M, Rodriguez H, Paulovich AG (2014) CPTAC assay portal: a repository of targeted proteomic assays. *Nat Methods* 11: 703 – 704

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A *et al* (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509: 582 – 587

Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, Houten SM, Amariuta T, Wolski W, Zamboni N, Aebersold R, Auwerx J (2016) Systems proteomics of liver mitochondria function. *Science* 352: aad0189

Wu JX, Song X, Pascovici D, Zaw T, Care N, Krisp C, Molloy MP (2016) SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Mol Cell Proteomics* 15: 2501 – 2514

Zhang Y, Bilbao A, Bruderer T, Luban J, Strambio-De-Castillia C, Lisacek F, Hopfgartner G, Varesio E (2015) The use of variable Q1 isolation windows improves selectivity in LC-SWATH-MS acquisition. *J Proteome Res* 14: 4359 – 4371

Zi J, Zhang S, Zhou R, Zhou B, Xu S, Hou G, Tan F, Wen B, Wang Q, Lin L, Liu S (2014) Expansion of the ion library for mining SWATH-MS data through fractionation proteomics. *Anal Chem* 86: 7242 – 7246

Zolg DP, Wilhelm M, Schnatbaum K, Zerweck J, Knaute T, Delanghe B, Bailey DJ, Gessulat S, Ehrlich HC, Weininger M, Yu P, Schlegl J, Kramer K, Schmidt T, Kusebauch U, Deutsch EW, Aebersold R, Moritz RL, Wenschuh H, Moehring T *et al* (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods* 14: 259 – 262

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res* 5: 2339 – 2347

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

nature communications

Article

https://doi.org/10.1038/s41467-024-49731-x

Sequence-to-sequence translation from mass spectra to peptides with a transformer model

Received: 16 January 2023

Accepted: 18 June 2024

Published online: 30 July 2024

Check for updates

Melih Yilmaz[1,5], William E. Fondrie[2,5], Wout Bittremieux[3,5], Carlo F. Melendez[4], Rowan Nelson[4], Varun Ananth[1], Sewoong Oh[1] & William Stafford Noble[1,4]

A fundamental challenge in mass spectrometry-based proteomics is the identification of the peptide that generated each acquired tandem mass spectrum. Approaches that leverage known peptide sequence databases cannot detect unexpected peptides and can be impractical or impossible to apply in some settings. Thus, the ability to assign peptide sequences to tandem mass spectra without prior information—de novo peptide sequencing—is valuable for tasks including antibody sequencing, immunopeptidomics, and metaproteomics. Although many methods have been developed to address this problem, it remains an outstanding challenge in part due to the difficulty of modeling the irregular data structure of tandem mass spectra. Here, we describe Casanovo, a machine learning model that uses a transformer neural network architecture to translate the sequence of peaks in a tandem mass spectrum into the sequence of amino acids that comprise the generating peptide. We train a Casanovo model from 30 million labeled spectra and demonstrate that the model outperforms several state-of-the-art methods on a cross-species benchmark dataset. We also develop a version of Casanovo that is fine-tuned for non-enzymatic peptides. Finally, we demonstrate that Casanovo's superior performance improves the analysis of immunopeptidomics and metaproteomics experiments and allows us to delve deeper into the dark proteome.

Mass spectrometry is currently the most popular analytical technique to characterize the proteome, by identifying and quantifying proteins present in complex biological systems[1]. During a bottom-up mass spectrometry proteomics experiment, proteins from a biological sample are enzymatically digested into peptides, their intact mass and charge are measured, and they are fragmented using tandem mass spectrometry. The fundamental challenge of mass spectrometry proteomics is then to determine the amino acid sequences of the resulting tandem mass (MS/MS) spectra. The standard approach to solve this spectrum identification problem is sequence database searching, during which peptides are simulated in silico using known digestion rules from a database of protein sequences potentially present in the biological samples, typically from a reference proteome. Next, each observed MS/MS spectrum is scored against a list of candidate peptides based on simplified peptide fragmentation rules, and the best-scoring peptide-spectrum match (PSM) is reported. Pioneered by the SEQUEST algorithm[2], dozens of database search engines have been subsequently developed and are very widely deployed[3].

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA. [2]Talus Bioscience, Seattle, USA. [3]Department of Computer Science, University of Antwerp, Antwerp, Belgium. [4]Department of Genome Sciences, University of Washington, Seattle, USA. [5]These authors contributed equally: Melih Yilmaz, William E. Fondrie, Wout Bittremieux. e-mail: William-noble@uw.edu

However, a fundamental requirement for sequence database searching is that the set of proteins that may be present in the sample is known in advance. While this is often the case for samples generated from species with well-characterized genomes, relying on a database prevents the detection of unexpected peptides. Such unexpected peptides include not just peptides derived from contaminant proteins or that arise due to variability in sample processing[4], but also biologically and clinically relevant peptides, such as peptides that deviate from the reference proteome due to genetic variation, peptides with unexpected post-translational modifications (PTMs), and peptides originating from foreign sources, such as microbes or consumed foods. Furthermore, there are tasks where generating a peptide database can prove impractical or even impossible. For example, the antigenic peptides presented by major histocompatibility complex (MHC) proteins—the "immunopeptidome"—are often generated from their parent proteins in an unpredictable manner, requiring at minimum every possible protein subsequence to be considered[5–7].

Similarly, constructing a peptide database for antibody sequencing is nearly impossible, due to the sequence variants created by V(D)J recombination[8]. Finally, creating an accurate database for mixtures of many organisms—metaproteomics—such as from microbiome or environmental samples, is often not feasible[9].

Such applications require the ability to sequence peptides directly from the acquired MS/MS spectra de novo. Early de novo peptide sequencing algorithms used heuristic search[10] or dynamic programming[11,12] to propose peptides for the observed MS/MS spectra. In addition to dynamic programming, the PepNovo algorithm[13] attempted to account for rules governing peptide fragmentation in its probabilistic score function and is closely related to the hidden Markov model employed by Fischer et al.[14] In 2015, the Novor algorithm[15] improved the state of the art by using a decision tree as the score function for its dynamic programming algorithm.

More recently, as in many other fields, deep learning has become the preferred solution for de novo peptide sequencing. DeepNovo[8] combines a convolutional neural network and a recurrent neural network to predict the subsequent amino acid when provided an MS/MS spectrum and a peptide prefix. SMSNet[16] uses a similar network architecture but reconciles the predicted sequences against a user-supplied peptide database. PointNovo[17], the successor to DeepNovo, leverages an order-invariant network architecture to model high-resolution MS/MS spectra[18]. Finally, pNovo 3[19] first generates candidate peptides for each MS/MS spectrum using dynamic programming, after which a final score is determined by matching the spectrum against a theoretical spectrum for each candidate peptide, simulated using the pDeep[20] learning-to-rank framework.

Despite the advances of these deep learning-based methods for de novo peptide sequencing, they still suffer from several limitations. De novo tools typically can only annotate a minority of MS/MS spectra compared to sequence database searching, they struggle with natively encoding high-resolution MS/MS data, and they employ complex neural network architectures and post-processing steps.

To address these issues, here we describe Casanovo, which reframes the de novo peptide sequencing task as a machine translation problem: like translating a sequence of words in a sentence from one language to another, Casanovo translates a sequence of peaks in an MS/MS spectrum into a sequence of amino acids of the generating peptide. To do so, we leverage the state-of-the-art architecture for modeling natural language—the transformer[21]. The transformer architecture allows Casanovo to directly use the *m/z* and intensity value pairs that comprise an MS/MS spectrum without discretization of the *m/z* axis and to directly output a predicted peptide sequence without a complicated dynamic programming step. We have previously[22] trained Casanovo on a limited collection of mass spectra from the multi-species benchmark used by Tran et al.[8] In this work, we present significant improvements to Casanovo and demonstrate its effectiveness at tackling common challenges with de novo peptide sequencing. We expand our training set to use 30 million confident PSMs from the MassIVE-KB spectral library[23], and we add a beam search decoding procedure to predict the best peptide for each MS/MS spectrum. We demonstrate that, together, these updates significantly improve Casanovo's already state-of-the-art performance. Additionally, we fine-tune a non-enzymatic version of Casanovo for tasks such as immunopeptidomics. We demonstrate how high-performance de novo peptide sequencing using Casanovo enables fast and effective immunopeptidome analysis, bolsters the characterization of metaproteomes, and sheds light on the dark proteome.

Results

A transformer architecture enables processing of raw mass spectra

Casanovo uses a transformer architecture to perform a sequence-to-sequence modeling task, from MS/MS spectrum to the generating peptide (Fig. 1). Transformers are built upon the attention function[21], which allows transformer models to contextualize the elements of a sequence; transformer models thus learn the relationships of sequence elements to one another and how their interactions should be interpreted. As such, the transformer architecture has found success in not only natural language processing, but also applications to biological sequences[24,25].

In Casanovo, each peak in an observed MS/MS spectrum is considered as an element in a variable length sequence. The *m/z* and intensity values of each peak are encoded using, respectively, a collection of sinusoidal functions and a learned linear layer, and these encodings are summed. The encoded peaks are then input into the transformer encoder, where context is learned between pairs of peaks in the MS/MS spectrum. The contextualized peak encodings are then used as input to the transformer decoder for predicting the peptide sequence.

The process of decoding proceeds in an iterative, autoregressive manner. We begin by providing the mass and charge of the observed precursor. The transformer decoder uses the contextualized peak encodings and the precursor information to begin predicting amino acids of the peptide. With the first predicted amino acid, we retain the top *k* residues, where *k* is a user-selected value for the number of

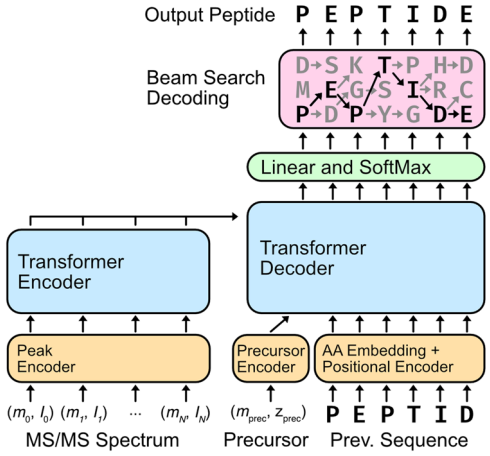


Fig. 1 | Casanovo performs de novo peptide sequencing using a transformer architecture. The peaks from each MS/MS spectrum are contextualized by the transformer encoder. The resulting peak encodings are then fed into the transformer decoder along with the observed precursor mass and charge to iteratively decode the peptide sequence. Casanovo uses a beam search decoding strategy, following the most promising sequence predictions until they terminate or exceed the precursor mass. The highest scoring peptide sequence is returned as the putative peptide that generated the MS/MS spectrum.

beams in our beam search. In each subsequent iteration, amino acids are added to the decoded peptide sequence, retaining the top k sequences until the decoded sequences for all of the beams have terminated or exceeded the precursor mass. Finally, the sequence with the highest score is retained as the putative peptide that generated the provided MS/MS spectrum.

In generating its predictions, Casanovo will inevitably fail to generate plausible peptides for some MS/MS spectra. For example, some MS/MS spectra contain too few fragment ions to be sequenced reliably, or the true generating peptide may bear a modification that is unknown to Casanovo. We therefore refine the PSMs proposed by Casanovo using a simple precursor mass filter: any PSMs for which the m/z of the peptide falls outside the specified tolerance of the observed precursor, including potential isotopes, is discarded. This filter eliminates many poorly scored peptides from consideration. In our evaluations, PSMs that would normally be removed by this filter were retained and ranked last among all PSMs assigned by Casanovo.

Casanovo outperforms state-of-the-art methods

To evaluate Casanovo, we first used the nine-species benchmark dataset originally created by Tran et al.[8] to compare the performance of four de novo peptide sequencing algorithms: Novor, DeepNovo, PointNovo, and Casanovo. For these comparisons, we used the publicly available, pretrained version of Novor to sequence the MS/MS spectra in the benchmark dataset. DeepNovo, PointNovo and Casanovo were trained in a cross-validated fashion, systematically training on eight species and testing on the remaining species. For DeepNovo, we used the models trained and provided by Tran et al.[8] for each of the cross-validation splits. For PointNovo, we cross-validated nine models from scratch using the code and settings provided by Qiao et al.[17]. This benchmark version of Casanovo, Casanovo$_{bm}$, employs a simple greedy decoding algorithm, rather than beam-search decoding. The results (Fig. 2a) revealed that Casanovo$_{bm}$ substantially improved peptide-level sequencing performance over Novor, DeepNovo and PointNovo, with an average precision of 0.81 for Casaonovo$_{bm}$ compared to 0.58, 0.70 and 0.74 for Novor, DeepNovo and PointNovo, respectively. These results are consistent across all nine species in the benchmark dataset (Supplementary Fig. S1).

We hypothesized that Casanovo could achieve even better performance if provided with a larger training set of higher quality PSMs; hence, we turned to the MassIVE-KB spectral library of human MS/MS proteomics data[23]. MassIVE-KB provided us with a set of 30 million high confidence PSMs, which we previously collected for training our GLEAMS embedding model[26]. This dataset contains not only a greater diversity of peptides and MS/MS spectra generated from multiple instruments, but also additional types of post-translational modifications. We therefore created a new version of the nine-species benchmark dataset using the same nine PRIDE datasets but including seven different types of variable modifications (methionine oxidation, asparagine deamidation, glutamine deamidation, N-terminal acetylation, N-terminal carbamylation, N-terminal NH_3 loss, and the combination of N-terminal carbamylation and NH_3 loss). In the process, we also fixed several problems that we uncovered in the previous benchmark, including adding consideration of isotope errors and eliminating peptides that occur in multiple species (see Methods for details). The final, revised benchmark dataset consists of 2.8 million PSMs drawn from 343 RAW files.

The results from evaluating with respect to this revised benchmark demonstrate the value of training from a much larger collection of higher quality PSMs (Fig. 2b). When trained on the MassIVE-KB dataset, the average precision of Casanovo increases from 0.83 to 0.95. Furthermore, Casanovo succeeds in making a larger proportion of predictions with m/z values that fall within 30 ppm of the observed precursor (signified by the location of the diamonds in Fig. 2b), increasing from 70% to 97%. Additionally, an analysis of spectrum identifications for all de novo sequencing tools on the nine-species benchmark dataset shows that correct Casanovo PSMs include almost all correct identifications of the competing de novo sequencing methods, as well as approximately 50% more correct PSMs that are unique to Casanovo (Supplementary Fig. S8). This version of Casanovo incorporates beam-search decoding, which improves both average precision and coverage compared to greedy decoding for the same model (Supplementary Fig. S3).

In one sense, this comparison is unfair because some of the spectra in the new version of the benchmark contain PTMs that cannot be identified by some of the competing methods. We therefore eliminated these spectra from each test set and then re-computed the precision-coverage curve. The results (Supplementary Fig. S4) are largely unchanged, suggesting that the PTMs contribute little to the observed overall differences in performance.

To better understand why the model trained on MassIVE-KB outperforms the one trained on the nine-species benchmark, we

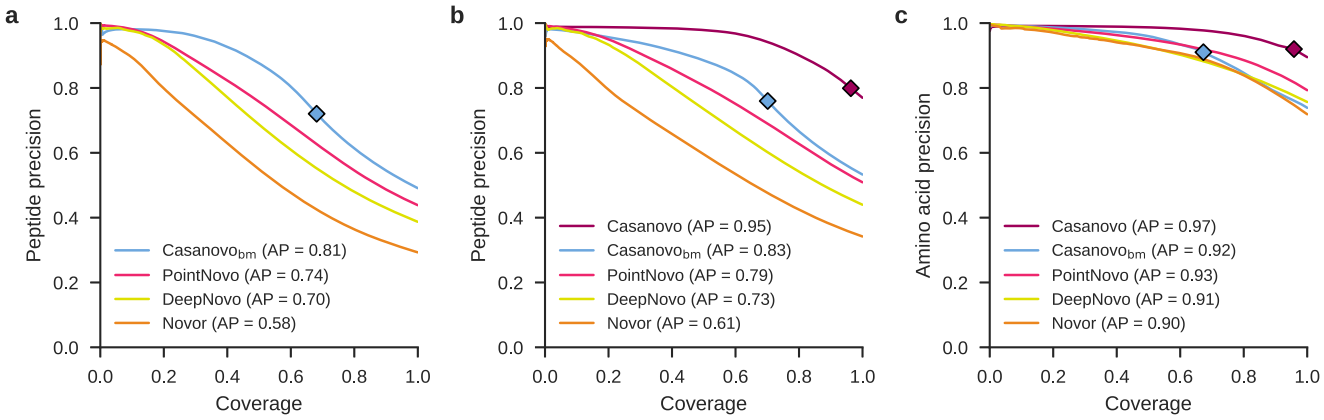


Fig. 2 | Casanovo outperforms PointNovo, DeepNovo, and Novor on a nine-species benchmark. a Casanovo maintains high peptide-level precision (the proportion of correctly predicted peptides) across all values of coverage (the proportion of spectra for which a prediction is made). Each curve is computed by sorting predicted peptides for all nine species according to their peptide-level confidence scores. For Casanovo, all peptides that pass the precursor m/z filter are ranked above peptides that do not pass the filter, and the boundary is indicated by a diamond on each curve. Average precision (AP) corresponds to the area under the precision-coverage curve. **b** Same as **a**, but using the revised benchmark and including a version of Casanovo trained on MassIVE-KB. **c** Casanovo's amino acid-level precision is greatly improved by the expanded training data provided by MassIVE-KB. The test set is the revised nine-species benchmark, with PSMs only containing modifications considered by both DeepNovo and Casanovo.

performed two follow-up experiments. First, we trained a series of Casanovo models on randomly sampled nested subsets of MassIVE-KB, ranging from 250,000 spectra to the full dataset of 28 million spectra. Each model was then evaluated with respect to the revised nine-species benchmark. The resulting learning curve (Supplementary Fig. S5) shows that the test set performance depends strongly on the size of the training set, though with diminishing returns after a million or so PSMs. Second, we directly compared a Casanovo model trained from a downsampled MassIVE-KB dataset to Casanovo$_{bm}$ which averages 9 models cross-validated on the nine-species benchmark, where the training sets contain approximately the same number of peptides (239,697 for Massive-KB and 246,713 for Casanovo$_{bm}$). We then evaluated both models using the revised nine-species benchmark. The results (Supplementary Fig. S6) show that the model trained from MassIVE-KB substantially outperforms Casanovo$_{bm}$, with the average precision increasing from 0.83 to 0.90. Thus, these results suggest that the improved performance of the MassIVE-KB model stems primarily from the improved quality of the data rather than the size of the data set.

In addition to evaluating Casanovo's ability to correctly predict whole peptides, we also evaluated Casanovo's ability to predict the individual amino acids of each peptide. We did so by ranking amino acids by their associated confidence score and then plotting a precision-coverage curve. We compared two versions of Casanovo (trained from the first benchmark and from MassIVE-KB) with DeepNovo and PointNovo on the revised nine-species benchmark with new modifications eliminated (Fig. 2c). The amino acid-level performance was consistent with the trends we observed in peptide-level performance, with Casanovo outperforming Novor, DeepNovo and PointNovo: Casanovo trained on MassIVE-KB achieves a remarkable average precision of 0.98.

Finally, to characterize the improved de novo sequencing performance of Casanovo across generating peptides of different lengths and precursor charge states, we compare all methods on subsets of the revised nine-species benchmark dataset. First, we divide spectra into three groups by charge state where groups contain precursors with 2+, 3+ and 4+ or higher charge states each, and plot peptide precision-coverage curves for each group (Supplementary Fig. S9). As expected, average precision is lower across all methods for groups with higher precursor charge states since those spectra tend to have more complex fragmentation patterns. However, the drop in performance is only 12% for Casanovo in precursors with 4+ or higher charge states versus precursors with 2+ charge states, thanks to the diversity of precursor charge states in its training data where 11% of precursors have 4+ or higher, whereas average precision for all competing methods decreases by more than 60%. Second, we bin spectra according to the length of their generating peptides into groups of short (fewer than 13 amino acids), medium (between 13 and 18 amino acids), and long (greater than 18 amino acids) peptides, and compare de novo sequencing performance in each group (Supplementary Fig. S10). Performance degrades for longer peptides because incorrect amino acid predictions tend to accumulate during decoding, but the observed decrease in average precision for Casanovo is much smaller relative to other methods, highlighting Casanovo's ability to accurately sequence long peptides as a key contributor to its improved performance.

Casanovo unravels the immunopeptidome

One important application of de novo peptide sequencing is the characterization of the peptides presented by major histocompatibility complexes (MHCs), which is commonly referred to as the "immunopeptidome." These antigen peptides are presented on the extracellular surface and serve as targets for immune cell recognition. However, because these antigen peptides are generated through lysosomal or proteasomal degradation, they do not exhibit the characteristic tryptic termini from most proteomics experiments. Consequently, the peptide search space is exponentially larger than considering only tryptic peptides—every peptide subsequence in a protein within a defined peptide length must be considered. Furthermore, mutations in these peptides are of particular interest, because these mutation-containing neoantigens may serve as tumor-specific markers to activate T cells and initiate antitumor immune responses. Unfortunately, expanding the search space to consider all possible mutated peptides is prohibitive both in terms of search speed and statistical power for traditional proteomics search engines.

Although immunopeptidomics is a prime application for de novo sequencing, naively applying Casanovo directly to immunopeptidomics data is problematic: the standard Casanovo model is heavily biased to predict tryptic peptides due to their overrepresentation in MassIVE-KB. To demonstrate this effect, we analyzed five mass spectrometry runs generated from MHC class I peptides isolated from MDA-MB-231 breast cancer cells[5] in two different ways: first using Casanovo and second by searching against a non-enzymatic digestion of the human proteome (see Methods). Among the peptides accepted at 1% FDR by the database search procedure, we observed a low proportion of "tryptic" peptides, i.e., peptides with C-terminal amino acids of K (1.12%) or R (0.80%). In contrast, among the top-scoring 10% of the Casanovo predictions, we observed a greater than six-fold increase in the rate of tryptic peptide predictions (5.87% K and 6.76% R).

We hypothesized that we could reduce this tryptic bias and produce a version of Casanovo that is better suited to immunopeptidomics data by fine tuning our existing model using data that lacks a tryptic bias. To create such a dataset, we combined data from two sources. First, we segregated PSMs from MassIVE-KB according to their C-terminal amino acid and then uniformly sampled up to 50,000 peptides within each group. For most amino acids, MassIVE-KB contained fewer than 50,000 PSMs, so for these we supplemented by randomly extracting additional PSMs from the PROSPECT collection[27] (Supplementary Table S1). We then split this new collection of 1 million PSMs into training, validation, and testing sets. We then fine-tuned our existing Casanovo model by training it until convergence on this non-enzymatic training set.

The resulting model, Casanovo$_{ne}$, performs markedly better than the original Casanovo model at predicting peptides in our held-out, non-enzymatic test set. On the held-out test set of 100,000 non-enzymatic PSMs, Casanovo$_{ne}$ achieves an average precision of 0.83, compared with 0.60 for the original Casanovo model on the same data (Fig. 3a). The predicted C-terminal amino acid frequencies are also much closer to the true frequencies, with K and R dropping to 1.81% and 1.79%, respectively (Fig. 3b).

We next used Casanovo$_{ne}$ to sequence the immunopeptidome of MDA-MB-231 breast cancer cells[5]. For each peptide predicted by Casanovo, we investigated whether it (1) occurs anywhere within the human proteome, and (2) occurs within the set of peptides detected using a database search procedure. We first searched the data against a non-enzymatic digestion of the human proteome using the Tide search engine[28] followed by Percolator post-processing[29], using settings similar to those in the original study[5] (see Methods). Out of 26,377 unique peptides predicted by Casanovo, 2459 match to the human proteome, and a majority of these overlap with the 1544 unique peptides identified by Tide at 1% FDR (Supplementary Fig. S11). Notably, these overlapping peptides are predicted with high confidence by Casanovo, almost all within the first 10,000 Casanovo predictions (Fig. 3c). Casanovo predicts an additional 1148 peptides that match to the human proteome but are not identified by Tide at 1% FDR, and further analysis shows that 751 (65.4%) of these peptides correspond to Tide hits that were not accepted at the 1% FDR threshold. To further investigate the plausibility of Casanovo predicted peptides as MHC antigens, we used NetMHCpan-4.1[30] to predict MHC binding affinity for these peptides. First, we

https://doi.org/10.1038/s41467-024-49731-x

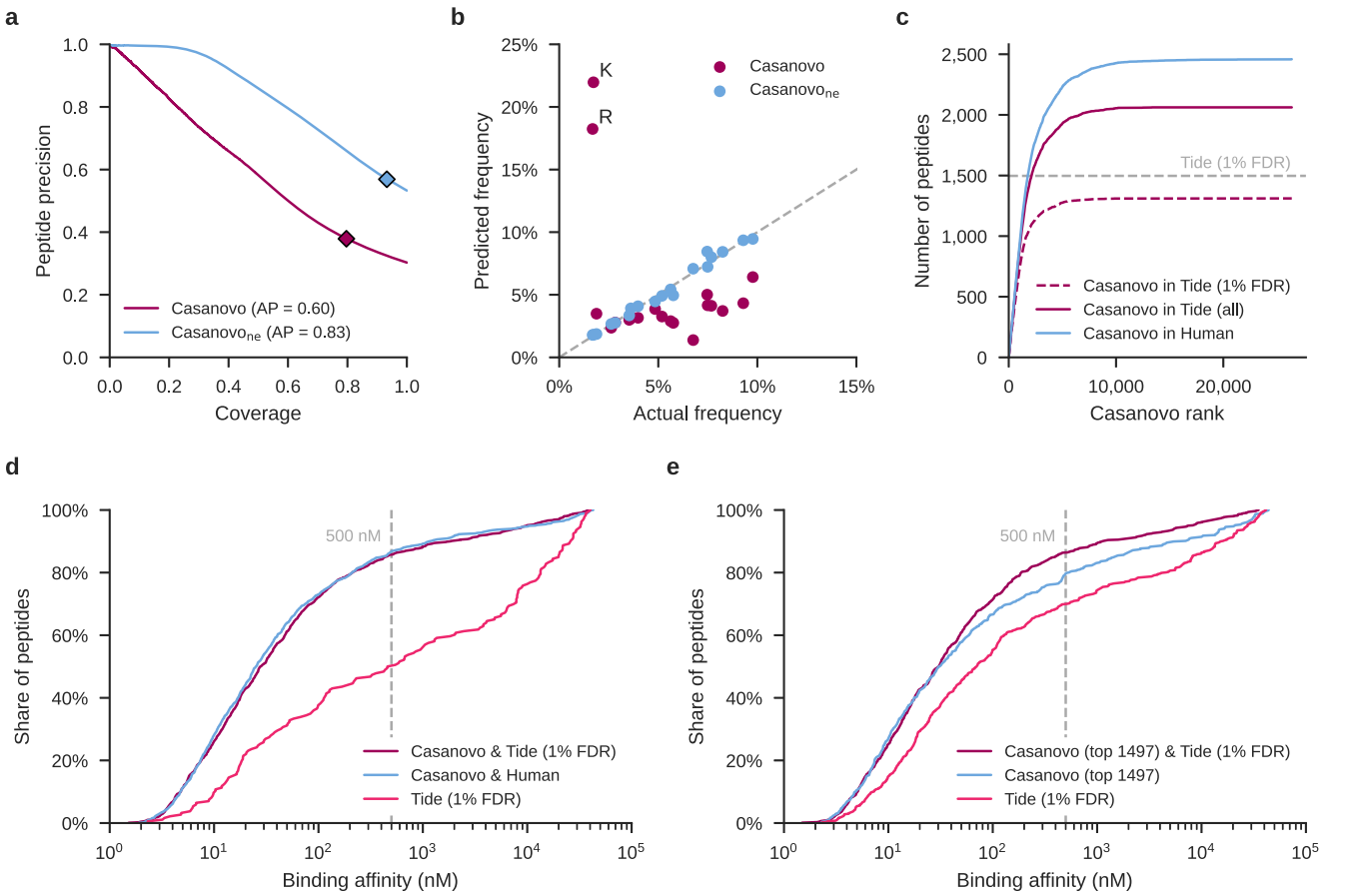


Fig. 3 | Fine-tuning reduces Casanovo's bias for tryptic peptides. a Fine-tuning Casanovo (Casanovo$_{ne}$) improves peptide-level precision on sequencing MS/MS spectra generated by non-tryptic peptides. **b** Casanovo$_{ne}$ predicts non-tryptic C-terminal peptides more readily than the standard Casanovo model, improving performance on the non-enzymatic validation set. **c** Casanovo detects many peptides that are present in the human proteome but are not detected via database search. The dashed dark pink line only includes peptides detected by database search within the 1% FDR threshold, whereas the solid dark pink line includes all peptides from the database search, irrespective of FDR threshold. **d** The peptides proposed by Casanovo generally have higher predicted binding affinities for the MHC class I receptor, matching the performance of a Tide database search. The vertical bar corresponds to the 500 nM binding affinity below which peptides are predicted to be MHC binders. **e** Similar to **d**, but considering only the 1497 peptides that are accepted at 1% FDR by Tide which yielded valid binding affinity predictions from NetMHCpan and a corresponding set of 1497 highest-confidence Casanovo peptides.

compared peptides that were identified by both Casanovo and database search with peptides that were predicted only by Casanovo and match to the human proteome. These two groups exhibit similar distributions of predicted binding affinity profiles, with 87% of peptides identified by Casanovo alone and 86% of those identified by both methods predicted to be MHC binders at 500 nM (Fig. 3d). In contrast, when we evaluate peptides that are identified by database search but not by Casanovo, the proportion of predicted MHC binders drops substantially to 50%. Overall, these results suggest that Casanovo not only identifies more peptides matching to the human proteome than the standard database search procedure, but the peptides Casanovo predicts are also more likely to bind MHC antigens.

We also explored an alternative method for comparing Casanovo and Tide results, which does not rely on mapping Casanovo predictions to the reference proteome. For this analysis, we consider the 1497 peptides identified by Tide at 1% FDR which yielded valid binding affinity predictions from NetMHCpan alongside the top 1497 highest confidence Casanovo predictions. The results (Fig. 3e) agree with those in Fig. 3d: the 960 peptides in common between the two sets of peptides achieve the highest proportion of MHC binders (86%), the Casanovo-only predictions achieve a slightly lower percentage (80%), and the Tide-only predictions have the lowest percentage of MHC binders (70%).

Casanovo accurately sequences peptides from complex metaproteomes

Proteomics applications extend far beyond the analysis of single model organisms or well-characterized biological systems. Indeed, there is growing interest in using mass spectrometry proteomics methods to investigate the dynamics of complex biological ecosystems—whether microbiomes or environmental specimens—for which the identities of its members cannot be known a priori. Due to the unknown complexity of the sample and even the lack of reliable reference proteomes for the likely species in the sample, these metaproteomics experiments are difficult to analyze. One solution to these problems is to search the spectra against a large database, such as one containing all the microbial sequences in public databases for a sample that is likely dominated by unknown microbes. This "big database" approach is widely used but suffers from a significant loss in statistical power due to the implicit multiple hypothesis testing correction that must be made to account for the size of the database. An alternative solution involves first subjecting the sample to genome sequencing, and then using the inferred peptide sequences as the basis for a "metapeptide" database. This approach yields better power to detect peptides[31] but requires the availability of a matched DNA sample and the additional cost associated with DNA sequencing.

We hypothesized that Casanovo's improved de novo sequencing capabilities would be useful in both scenarios—either in the presence

https://doi.org/10.1038/s41467-024-49731-x

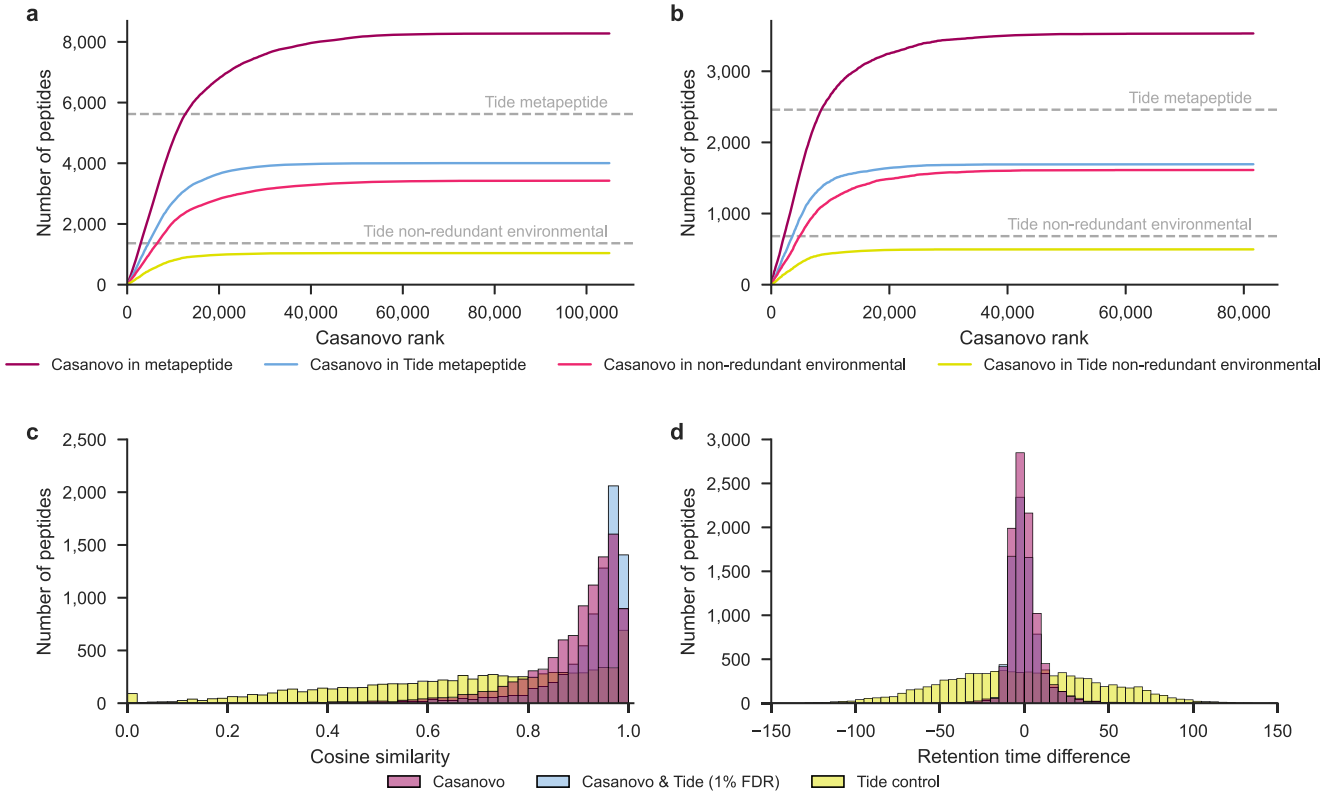


Fig. 4 | Casanovo improves power to detect peptides from metaproteomics samples. Casanovo assigns more peptides matching the metapeptide database and the non-redundant environment database than Tide and Percolator at 1% FDR in seawater samples from **a** the Bering Sea and **b** the Chukchi Sea. Peptides are ranked according to the Casanovo confidence score, assigning each peptide the maximum score across all three runs from each sampling location. Horizontal lines indicate the total number of distinct peptides detected by Tide+Percolator, searching against two different databases. **c** The PSMs assigned by Casanovo at a 1% error rate and Tide and Percolator at 1% FDR have high cosine similarities to the predicted MS/MS spectra for the respective peptides from Prosit when compared to control PSMs sampled from the Tide search results with > 10% FDR. Each group represents the aggregated results for Bering and Chukchi Sea data, as well as non-redundant environmental and metapeptide databases. **d** The PSMs assigned by Casanovo at a 1% error rate and Tide and Percolator at 1% FDR closely align with the predicted retention times from Prosit.

or absence of a metapeptide database. To test this hypothesis, we applied Casanovo to data from six previously published ocean metaproteomics samples, three from the Bering Strait and three from the Chukchi Sea[31]. Critically, these samples were also subjected to DNA sequencing; hence, in addition to the non-redundant environmental database, we also have a metapeptide database for each sampling location.

We began by measuring the extent to which peptides detected by Casanovo occur within the corresponding metapeptide database or within the larger, non-redundant protein database. Because these samples were digested using trypsin, we used the standard Casanovo model, trained from the tryptic MassIVE-KB dataset. To control the error rate for the matching of Casanovo predictions to these databases, we employed a procedure similar to target-decoy competition used in the false discovery rate estimation for database search (see Methods for details) and only considered as correct Casanovo peptides found in the corresponding database that fall within the 1% random matching threshold. Using this logic, we obtain much better power to detect peptides using Casanovo than using a standard database search procedure against the metapeptide database (Fig. 4a–b). In particular, when we search the data against the metapeptide database using Tide followed by Percolator[29], we detect 5623 peptides at 1% FDR in the Bering Strait data and 2460 peptides in the Chukchi Sea data. In contrast, if we run Casanovo and accept as correct only peptides that appear in the metapeptide database (subject to our 1% random matching criterion), then we detect 8277 and 3532 peptides, respectively, in the two datasets, representing increases of 47% and 44%. Casanovo also outperforms database search when we consider the non-redundant protein database rather than the sample-specific metapeptide database. We detect 1364 peptides in the Bering Strait data and 682 peptides in the Chukchi Sea data at 1% FDR by searching the non-redundant environmental database using Tide and Percolator. In comparison, Casanovo predictions, filtered at 1% error rate using the environmental database, detect 3425 and 1612 peptides, respectively, representing increases of 151% and 136%, respectively.

When using both metapeptide databases or the non-redundant environment database, Casanovo detects most of the peptides identified by Tide database search and Percolator, where it respectively detects 71% and 75% of Tide identifications on metapeptide and non-redundant databases, while also detecting a substantial number of additional unique peptides (Supplementary Fig. S12).

To validate the peptides that were detected by Casanovo but not the database search, we used the Prosit machine learning tool[32] to predict spectrum peak intensities and retention times for peptide identifications. First, we compared the cosine similarities between the observed and predicted MS/MS spectrum peak intensities across three groups of peptides: peptides only predicted by Casanovo that matched to the database with 1% error, peptides detected both by Casanovo and by Tide and Percolator at 1% FDR, and a control group of peptides detected by Tide and Percolator with >10% FDR. The control group was randomly sampled to be the same size as the Casanovo-only group. The results (Fig. 4c) indicate that the Casanovo-only identifications have a high concentration of high cosine similarity peptides, similar to the overlapping identifications between Casanovo and

database search. This stands in contrast with the control group, which exhibits a much broader distribution of cosine similarities.

Second, we compared the observed retention times with the predicted retention times from Prosit for the same three groups of peptides. For each group, we calibrated the predicted retentions times to the observed retention times using linear regression (Supplementary Fig. S13). We observed that the peptides detected only by Casanovo and those detected by Casanovo and Tide had a similar slope and resulted in similar residual distributions (Fig. 4d). When compared against the control group, the residual distributions for peptides only detected by Casanovo and those detected by Casanovo and Tide are close to zero.

Ultimately, Casanovo does not yet allow us to achieve as much power with the non-redundant database as with the metapeptide database. For example, for the Bering Strait data, the union of the 3425 peptides detected using Casanovo and the 1364 peptides detected using database search is 3750, which is fewer than the 5623 peptides detected using the metapeptide database. (The corresponding numbers for the Chukchi Sea data are 1798 and 2460.) This difference is perhaps not surprising, because the environmental non-redundant database is incomplete: 3715 of the 5623 peptides found by the database search procedure in the Bering Strait metapeptide database are not even present in the environmental database. Thus, a rigorous FDR control procedure for de novo peptide sequencing is needed in order to rescue the many peptides that are correctly detected by Casanovo but cannot be validated by matching to a database.

Casanovo shines a light on the dark proteome

The "dark matter" of mass spectrometry-based proteomics consists of MS/MS spectra that are observed repeatedly across experiments but consistently fail to be identified. In many cases, these MS/MS spectra may have been generated by peptides that are not in the canonical human proteome, because they represent contaminant peptides, result from non-standard enzymatic cleavage, or contain sequence variants. We hypothesized that Casanovo could shed light on some of this dark matter.

Accordingly, we applied Casanovo to a collection of MS/MS spectra drawn from a previous analysis[26], in which 511 million human spectra from MassIVE were grouped into 60 million clusters, and the clusters were systematically analyzed using targeted open modification searching of representative spectra. The analysis yielded a collection of 39 million unidentified clusters, containing a total of 207 million MS/MS spectra. For our analysis, we selected 3.4 million of these unidentified, clustered MS/MS spectra from eight randomly selected MassIVE datasets. These MS/MS spectra belong to 573,597 distinct clusters. Because we were investigating spectra that had already failed to be identified using a standard, tryptic pipeline, we opted to use the non-enzymatic Casanovo model (Casanovo$_{ne}$) to assign a peptide to each selected MS/MS spectrum, eliminating peptides for which the predicted m/z falls outside the associated mass range. This analysis yielded a total of 1.3 million predicted peptides.

We sought to ascertain how well Casanovo had assigned peptide sequences to these dark matter clusters by addressing this question in two complementary ways. First, we identified all clusters in which a plurality (and at least two) of the spectra were assigned to the same peptide sequence, and then we mapped those peptides to the human reference proteome, allowing at most one amino acid mismatch. The first step of this procedure assigns peptides to 89,250 (16%) of the clusters, of which 65% could be matched to the human proteome. The clusters identified in this fashion vary in size, ranging from 2 to 542 spectra per cluster, but when we limited the above analysis only to clusters larger than a certain size, we observed that the shares of identified clusters more than doubled (Supplementary Fig. S14). Second, we performed a complementary analysis, first eliminating all predicted peptides that do not occur within the human proteome (again, allowing one mismatch) and then finding clusters with two or more spectra assigned the same sequence and no other spectrum assigned to a different sequence. This procedure assigns peptides to 52,523 clusters, corresponding to 9% of all previously unidentified clusters. The overlap between the two approaches—plurality vote followed by proteome matching or vice versa—is high: 98% of the 52,523 clusters overlapped with the clusters from the previous analysis. Overall, Casanovo is able to assign peptides to 196,724 of the 3.4 million unidentified MS/MS spectra using the combination of these two strategies.

One potential reason for an MS/MS spectrum to remain unidentified is the presence in the generating peptide sequence of a genetic variant that does not appear in the reference proteome. To investigate whether Casanovo is identifying such sequences, we looked more closely at the subset of Casanovo cluster assignments that match to the human proteome with a single amino acid mismatch, focusing on the 51,555 assignments that agree between the two methods described above. Two pieces of evidence suggests that these peptides are indeed enriched for genetic variants. First, we observe an enrichment for amino acid substitutions that can be explained by a corresponding single-nucleotide substitution. Among the Casanovo predictions, 83.4% correspond to a potential single-nucleotide substitution, compared with only 38.6% of all possible amino acid substitutions that fit this criterion. Second, we see a strong enrichment for substitutions with positive BLOSUM62 scores[33]. The BLOSUM score is an integerized log-odds score indicating the empirical substitutability of one amino acid for another. In the BLOSUM62 matrix, only 11% of the 380 non-diagonal entries are positive. However, if we rank the Casanovo-predicted substitutions by frequency, we find that the top five substitutions have BLOSUM scores of 1 or 2 (Supplementary Table S3). This observation strongly suggests that Casanovo is predicting substitutions that are biochemically plausible.

Discussion

Casanovo's excellent performance derives from two sources: the availability of a large, high-quality set of training data, and the use of a machine learning architecture that can make use of that data. Our experiments suggest that the carefully curated MassIVE-KB collection provides particularly good training data. This is likely because the dataset was derived from a massive collection of 669 million spectra, in combination with extremely stringent FDR control. In particular, the data were searched at 1 % FDR, after which only the top 100 PSMs for each unique precursor were retained, corresponding to 30 million high-quality PSMs (uniformly 0 % FDR from the original searches).

The transformer architecture is uniquely suited to contextualize the elements of variable length sequences and has therefore proven immensely successful in modeling natural language. In comparison to recurrent neural networks, the transformer architecture is able to learn long-range dependencies between the elements of a sequence and can be parallelized for efficient training. By encoding the peaks of a mass spectrum as a sequence, similar to tokenizing the words of a sentence, Casanovo leverages the strengths of the transformer architecture and the rapid advances pioneered in large language models to improve de novo peptide sequencing from MS/MS spectra[34]. One important open question, which we leave for future work, is how the number of model parameters impacts de novo sequencing performance.

Casanovo's utility extends beyond the applications we have demonstrated here. Most obviously, any application in which a peptide database is unavailable, incomplete, or extremely large may benefit from de novo sequencing, such as paleoproteomics, forensics, or astrobiology[35]. However, even in the analysis of human or model organism data, Casanovo can assist in the detection of "foreign" spectra, i.e., spectra generated by peptides that are not present in the database. Such foreign spectra might correspond to contaminants introduced during the experiment itself, but they can also represent

microbial species, genetic variation, or trans-spliced peptides. In general, we envision applying Casanovo as a post-processor for spectra that fail to be assigned a peptide during a standard database search procedure, akin to the last stage of a cascade search[36].

One important application of de novo sequencing that we have not yet explored is antibody sequencing. However, a recent publication carried out a systematic comparison of six de novo sequencing tools, including Casanovo, on the problem of antibody sequencing[37]. The results show that Casanovo strongly outperforms the competing methods by all of the measures that the authors considered. Notably, this comparison employed a version of Casanovo that used greedy decoding and was trained on only 2 million spectra. Hence, our results (Fig. 2b) suggest that the version of Casanovo trained from 30 million spectra will yield even better antibody sequencing performance.

We anticipate many opportunities for fine-tuning the Casanovo model for particular applications. Our analysis with the non-enzymatic model suggests that Casanovo's enzymatic bias can be adjusted by using a relatively small amount of training data. Thus, in the short term, we plan to train variants of Casanovo that are appropriate for a variety of different cleavage enzymes. The Casanovo software makes such fine-tuning straightforward, so any user interested in adapting the model to a particular experimental setting should be able to do so. Longer term, an ideal model would take as input a spectrum along with relevant metadata, such as the digestion enzyme, collision energy, and instrument type, and predict accurately for many different types of experimental settings.

The potential for deep learning methods to improve our ability to perform de novo sequencing has now been widely recognized. While this paper was under review, at least six additional deep learning de novo sequencing methods have been published, including GraphNovo[38], PepNet[39], Denovo-GCN[40], Spectralis[41], π-HelixNovo[42], and NovoB[43]. Clearly, the field would benefit from an exhaustive and rigorous benchmark comparison of this growing field of tools.

On a related note, at this stage one of the key bottlenecks in the field is the absence of a rigorous method for confidence estimation for de novo sequencing. In our metaproteomics analysis, we have matched Casanovo predictions to a target and corresponding decoy peptide database, but such an approach misses out on the power of de novo sequencing to assign peptides to foreign spectra. Thus, an open question is whether, for a given data-dependent acquisition dataset, Casanovo outperforms a standard database search procedure in terms of statistical power to detect peptides. Trained from sufficiently large training sets, we may be approaching the end of database searching as the go-to method for analysis of DDA tandem mass spectrometry data.

Methods

Casanovo

Casanovo consists of a transformer encoder and decoder stack as described by Vaswani et al.[21], which are respectively responsible for learning latent representations of the input spectrum peaks and decoding the amino acid sequence of the spectrum's generating peptide. The encoder takes d-dimensional spectrum peak embeddings as input and outputs d-dimensional latent representation vectors for each peak. Subsequently, the decoder takes as input these representations of prefix amino acids, coupled with a d-dimensional precursor embedding encapsulating precursor m/z and charge information, to predict the next amino acid in the peptide sequence. We discuss different aspects of our modeling strategy in detail below.

Spectrum preprocessing. We preprocess each mass spectrum by removing noise peaks and scaling the peak intensities before they are transformed into input embeddings for Casanovo. First, we discard any peaks outside the range 50–2500 m/z, as well as any peaks within 2 Da of the observed precursor mass. We then remove any peaks with an intensity value lower than 1% of the most intense peak's intensity, and we retain up to 150 of the most intense peaks in the spectrum. Finally, peaks intensities are square-root transformed and then normalized by dividing by the sum of the square-root intensities.

Input embeddings. Each spectrum $S = \{(m_j, I_j)\}_{j=1}^{N}$ is a bag of peaks, where each peak (m_j, I_j) is a 2-tuple representing the m/z value and intensity of the peak. For the task of de novo peptide sequencing, the most important relationships for our model to learn are how the spacing of m/z values between each pair of peaks corresponds to the peptide ions that may have generated them. Secondarily to the spacing of m/z values, the intensity of each peak also contains information about the generating ion; for example, y-ions are generally more intense than b-ions for some fragmentation methods. Given this prior knowledge, we chose embedding methods that would enable Casanovo to learn from the spacing of m/z values and that would emphasize the relative importance of these peak attributes for the de novo sequencing task.

We use a fixed, sinusoidal embedding[21] to project each m/z value to a d-dimensional vector, the m/z embedding f. Specifically, we create the m/z embedding from an equal number of sine and cosine waveforms spanning the wavelengths from 0.001 to 10,000 m/z, where each feature in the embedding f_i is a value from one waveform (Supplementary Fig. S15A). Let $\lambda_{\max}$ be the maximum wavelength, $\lambda_{\min}$ be the minimum wavelength, i be the index of the feature (zero-based), and d be the number of features. We begin by defining the number of features that are to be represented by sine and cosine waveforms as $d_{\sin}$ and $d_{\cos}$, respectively:

$$d_{\sin} = \left\lceil \frac{d}{2} \right\rceil \tag{1}$$

$$d_{\cos} = d - d_{\sin} \tag{2}$$

The encoded features are then calculated as:

$$f_i(m_j) = \begin{cases} \sin\left(m_j \Big/ \left(\frac{\lambda_{\min}}{2\pi}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{i/(d_{\sin}-1)}\right)\right), & \text{for } i \le d/2 \\ \cos\left(m_j \Big/ \left(\frac{\lambda_{\min}}{2\pi}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{(i-d_{\sin})/(d_{\cos}-1)}\right)\right), & \text{for } i > d/2 \end{cases} \tag{3}$$

where $\lambda_{\max} = 10{,}000$ and $\lambda_{\min} = 0.001$ in Casanovo.

These input embeddings provide a granular representation of high-precision m/z information and, similar to relative positions in the original transformer model[21], may help the model attend to m/z differences between peaks, which are critical for identification of amino acids in the peptide sequence. In the m/z embeddings, we chose high-frequency waveforms to capture the fine structure present in a mass spectrum, such as that introduced by isotopes and near isobaric species. The waveforms then capture more distant relationships as they progress to lower frequencies; thus, if one were to subtract the m/z of one peak from another, the features that are activated depend on the scale of their relationship. While consecutive b- and y-ions may activate one set of features in f, complementary b- and y-ions would likely activate a later set due to their larger m/z difference. Furthermore, the cosine similarity between pairs of m/z embeddings are negatively correlated with their m/z (Supplementary Fig. S15B). We postulate that this property preserves information about m/z distances—which are critical for de novo peptide sequencing—in a manner that is readily accessible to the subsequent transformer layers.

The intensity, which is measured with lower precision than the m/z value, is embedded by projection to d dimensions through a linear layer, after which the m/z and intensity embeddings are summed to produce the input peak embedding. However, in developing

Casanovo, we found that using a sinusoidal encoding for intensity as well as *m/z* performs similarly.

Precursor information, used as input to the decoder, consists of the total mass $m_{\text{prec}} \in R$ and charge state $c_{\text{prec}} \in \{1, \ldots, 10\}$ associated with the spectrum. We use the same sinusoidal position embedding as peak *m/z*'s for m_{prec}; c_{prec} is embedded using an embedding layer, and the embeddings are summed to obtain the input precursor embedding.

Modeling de novo peptide sequencing as a sequence-to-sequence task. The transformer architecture in Casanovo follows the standard encoder-decoder design of Vaswani et al.[21]. The process begins by embedding the peaks of a mass spectrum to obtain input embeddings f, which are then contextualized using the transformer encoder stack. Thus, a full mass spectrum consisting of k peaks is represented as an unordered sequence of peak embeddings $g \in \mathbb{R}^{k \times d}$. The self-attention mechanism of the transformer encoder learns relationships between these peaks and outputs a contextualized embedding of each peak in the mass spectrum $\hat{g} \in \mathbb{R}^{k \times d}$.

The decoding process begins by feeding both the precursor embedding and the contextualized spectrum embedding $\hat{g}$ into the transformer decoder stack. Decoding proceeds in an autoregressive manner; up to a maximum number of iterations, the decoder stack will attempt to predict the next amino acid in the generating peptide from C-terminus to N-terminus. During the decoding phase, a learned representation of the previously predicted amino acid is concatenated to the input into the decoder for each iteration and summed with a sinusoidal positional embedding of its position in the sequence. The output of the decoder are scores $s \in \mathbb{R}^{p \times v}$ representing how confident Casanovo is about each amino acid it has predicted for a peptide sequence of length p and an amino acid vocabulary of size v.

Training and inference strategy. Taking the previously described embeddings as input, the transformer outputs scores which are treated as a probability distribution over the amino acid vocabulary for the next position in the sequence at each decoding step. The amino acid vocabulary includes 20 canonical amino acids (with cysteine considered to be carbamidomethylated), post-translationally modified versions of three of them (oxidation of methionine and deamidation of asparagine or glutamine), N-terminal modifications (acetylation, carbamylation, loss of ammonia, and the combination of loss of ammonia and carbamylation), plus a special `stop` token to signal the end of decoding, yielding a total of 28 tokens. During training, the decoder is fed the amino acid prefix for the ground truth peptide following the teacher forcing paradigm[44]. Cross-entropy between the model output probabilities and a binary matrix representing the amino acid sequence of the ground truth peptide is minimized as the objective function. During inference, beam search is used to find the highest-scoring predicted peptide sequence, with k a user-specified value for the number of beams. At each prediction step, for every peptide prefix considered, the k top-scoring amino acids are selected, after which the k top-ranked amino acid sequences are used for the subsequent decoding step. Beams are terminated when the `stop` token is predicted, the predicted peptide mass is similar to (given the precursor mass tolerance) or exceeds the precursor mass, or the pre-defined maximum peptide length of $\ell = 100$ amino acids is reached. As final prediction, the top-scoring peptide sequence that fits the precursor mass tolerance (optionally accounting for isotope offsets) is selected. If no peptide prediction fits the precursor mass tolerance, the top-scoring peptide sequence with a non-matching peptide mass is selected.

Model and training hyperparameters. We train models with nine layers, embedding size $d = 512$, and eight attention heads, yielding a total of ~47M model parameters. A batch size of 32 spectra and 10^{-5} weight decay is used during training, with a peak learning rate of 5×10^{-4}. The learning rate is linearly increased from zero to its peak value in 100,000 warm-up steps, followed by a cosine shaped decay. The MassIVE-KB model was trained for a single epoch on 30 million PSMs from the MassIVE-KB dataset, which took approximately 8 days on 4 RTX 2080 Ti GPUs, while evaluating the performance on the validation set after every 50,000 iterations. The final model weights were taken from the snapshot with the lowest validation loss. This model was fine-tuned on the non-enzymatic training dataset using a peak learning rate of 5×10^{-5}. We selected the fine-tuned model with the minimal validation loss, which occurred after 5 epochs. These model hyperparameters—number of layers, embedding size, number of attention heads, and learning rate schedule—are used for all downstream experiments unless otherwise specified.

Precursor *m/z* filtering. A critical constraint in de novo peptide sequencing requires the difference between the total mass of the predicted peptide m_{pred} and the observed precursor mass m_{prec} of the spectrum to be smaller than a threshold value ϵ (specified in ppm) for the predicted sequence to be plausible: $\Delta m_{ppm} = \frac{|m_{\text{prec}} - m_{\text{pred}}| \times 10^6}{m_{\text{prec}}} < \epsilon$. Therefore, in addition to providing precursor information as an input for the model to learn from, we filter out peptide predictions that do not satisfy the above constraint. The threshold value ϵ is a property of the mass spectrometer that the data is collected with, and hence is known at inference time. Accordingly, we choose ϵ based on the precursor mass error tolerance used in the database search to obtain ground truth peptide sequences for the test data.

Datasets

MassIVE-KB dataset. A large-scale, heterogeneous dataset derived from the MassIVE knowledge base (MassIVE-KB; v.2018-06-15) was used to develop Casanovo[23]. The MassIVE-KB data set consists of 31 TB of human data from 227 public proteomics datasets, containing over 669 million MS/MS spectra. MassIVE-KB contains a designated subset of 30,506,973 "high quality" PSMs, identified by applying a strict (~0%) PSM-level FDR filter and then selecting at most 100 PSMs for each combination of peptide sequence and charge. These 30 million PSMs were randomly split so that the training, validation and test sets are disjoint at the peptide-level and consist of approximately 28 million training PSMs, 1 million validation PSMs, and 1 million test PSMs.

Nine-species benchmark dataset. We created a new version of the nine-species benchmark originally described by Tran et al.[8] To do so, we downloaded the RAW files from the same nine PRIDE projects (Supplementary Table S2) and converted them to MGF format using the ThermoRawFileParser v1.3.4. We also downloaded the corresponding nine Uniprot reference proteomes and constructed a Tide index for each one, using Crux version 4.1. For one species (*Vigna mungo*), no reference proteome is available, so we used the proteome of the closely related species *Vigna radiata*. We specified Cys carbamidomethylation as a static modification and allowed for the following variable modifications: Met oxidation, Asn deamidation, Gln deamidation, N-term acetylation, N-term carbamylation, N-term NH_3 loss, and the combination of N-term carbamylation and NH_3 loss by using the tide-index options –mods-spec 1M+15.994915, 1N+0.984016, 1Q+0.984016 –nterm-peptide-mods-spec 1X+42.010565, 1X+43.005814, 1X-17.026549, 1X+25.980265 –max-mods 3. Note that one of the nine experiments (*Mus musculus*) was performed using SILAC labeling, but we searched without SILAC modifications and hence include in the benchmark only PSMs from unlabeled peptides. Each index also contains a shuffled decoy peptide corresponding to each target peptide. Each MGF file was searched against the corresponding index using the precursor window size and fragment bin tolerance specified in the original study (Supplementary Table S2). We used XCorr scoring with Tailor calibration[45], and we allowed for 1 isotope error in the selection of

candidate peptides. All search results were then analyzed jointly per species using the Crux implementation of Percolator, with default parameters. For the benchmark, we retained all PSMs with Percolator q value < 0.01. We identified 13 MGF files with fewer than 100 accepted PSMs, and we eliminated all of these PSMs from the benchmark. We then post-processed the PSMs to eliminate peptides that are shared between species. Among the 229,984 unique peptides, we identified 3797 (1.7%) that occur in more than one species. For each such peptide, we selected one of the associated species at random and then eliminated all PSMs containing that peptide in other species. Note that when identifying shared peptides between species, we considered all modified forms of a given peptide sequence to be the same. Hence, if a given peptide appears in more than one species, then that peptide, including all its modified forms, is randomly assigned to a single species and eliminated from the others. The final benchmark dataset consists of 2.8 million PSMs drawn from 343 RAW files. The revised nine-species benchmark is available on MassIVE at https://doi.org/10.25345/C52V2CK8J.

Evaluation metrics

We use precision calculated at the amino acid and peptide levels[8,11,13] as a function of coverage over the test set as performance measures to evaluate the quality of a given model's predictions. In each case, for each spectrum we compare the predicted sequence to the ground truth peptide from the database search. Following Tran et al.[8], for the amino acid-level measures we first calculate the number N^a_{match} of matched amino acid predictions, defined as all predicted amino acids which (1) differ by < 0.1 Da in mass from the corresponding ground truth amino acid, and (2) have either a prefix or suffix that differs by no more than 0.5 Da in mass from the corresponding amino acid sequence in the ground truth peptide. We then define amino acid-level precision as $N^a_{\text{match}}/N^a_{\text{pred}}$, where N^a_{pred} is the number of predicted amino acids. For peptide predictions, a predicted peptide is considered a correct match if all of its amino acids are matched. Among a collection of N^p_{orig} spectra, if our model makes predictions on a subset of N^p_{pred} and correctly predicts N^p_{match} peptides, we define coverage as $N^p_{\text{pred}}/N^p_{\text{orig}}$ and peptide-level precision as $N^p_{\text{match}}/N^p_{\text{pred}}$. To plot a precision-coverage curve, we sort predictions by the confidence score provided by the model. Amino acid-level confidence scores are obtained by applying a softmax to the output of the transformer decoder, which is a proxy for the probability of each predicted amino acid to occur in the given position along the peptide sequence. Casanovo reports the mean probability score over all amino acids as a peptide-level confidence score, and reports a modified score at the amino acid level, computed as the mean of the peptide score and the individual amino acid probability score.

Competing methods

We downloaded DeepNovo weights from https://github.com/nh2tran/DeepNovo/tree/PNAS on Sep 6, 2022. Similar to Casanovo$_{bm}$, DeepNovo and PointNovo were trained in a cross-validated fashion using the original nine-species benchmark, systematically training on eight species and testing on the remaining species. Accordingly, nine different sets of pre-trained DeepNovo weights were available, and the corresponding set of weights were used for testing on each species data set. In the absence of pre-trained PointNovo weights, we cross-validated nine models from scratch by training on eight species and testing on the remaining species. We downloaded the PointNovo code provided by Qiao et al.[17] from https://zenodo.org/records/3960823 on Mar 27, 2023.

We downloaded Novor v1.05.0573 from https://github.com/compomics/searchgui/tree/master/resources/Novor on Dec 3, 2022.

Creating a non-enzymatic dataset

To create a dataset of PSMs that does not exhibit tryptic bias, we selected PSMs with a uniform distribution of amino acids at the C-terminal peptide positions from two datasets: MassIVE-KB[23] and PROSPECT[27]. The MassIVE-KB dataset contains 30 million PSMs and consists entirely of data generated using trypsin; hence, only a small proportion of the MassIVE-KB peptides do not end in K or R, corresponding to those that appear at the C-terminus of the corresponding protein. The PROSPECT dataset is a collection of 61 million PSMs generated from synthetic peptides. We performed three filtering steps on this dataset: (1) removed duplicate peaks with identical *m/z* values from each spectrum, (2) eliminated spectra with fewer than 20 peaks, and (3) eliminated spectra with Andromeda score less than 70, selecting the highest-scoring peptide for each spectrum. To avoid over-representation of some peptides in this dataset, we randomly selected at most 100 PSMs per unique peptide, similar to the processing that was done during the creation of the MassIVE-KB dataset. This preselection step reduced the size of the PROSPECT dataset to 12.6 million PSMs. Finally, to create a non-enzymatic dataset containing 1 million peptides, we selected 50,000 PSMs for each canonical amino acid. These PSMs were selected at random from MassIVE-KB, then supplemented as necessary from PROSPECT to obtain the desired count (Supplementary Table S1). This dataset contained PSMs from 247,859 unique peptides, which were randomly split into training, validation and test sets in the ratio 80/10/10. The non-enzymatic dataset with the training, validation and test splits is available on MassIVE at https://doi.org/10.25345/C5KS6JG0W.

Immunopeptidome analysis

For the analysis of the MHC class I peptides, we used the Tide search engine and adopted search settings from the original publication[5]. To create the peptide index, we ran tide-index allowing M oxidation or phosphorylation of S/T/Y, with a maximum of one modification per peptide. We set the digestion to be "non-specific-digest," allowed zero missed cleavages, and specified a peptide length range of 8–15 amino acids. Using the canonical human reference proteome downloaded from Uniprot on July 17, 2022, the resulting index contains 286,319,284 peptides and an equal number of shuffled decoy peptides. We searched the data using the tide-index command, specifying a precursor window size of 30 ppm and using Tailor calibration. The resulting sets of PSMs from all five runs were analyzed jointly using Percolator with default settings. All of the above commands were implemented within Crux[46] version 4.1-2fab3c91-2022-11-09. For comparative analysis between Casanovo predictions and database search results, only sequences within a peptide length range of 9–15 amino acids were considered. NetMHCpan-4.1 was used to predict MHC binding affinities for peptide sequences[30]. Binding affinity was predicted for 9-mer amino acid motifs in reference to the HLA-A02:01, HLA-A02:17, HLA-B41:01, HLA-B40:02, HLA-C02:02, and HLA-C17:01 alleles of the MHC molecule.

Metaproteomics analysis

We analyzed data from six mass spectrometry runs, three replicates each from the Bering Strait (BSt) and the Chukchi Sea (CS)[31], downloaded in mzXML format from https://noble.gs.washington.edu/proj/metapeptide. From the same URL, we also downloaded the two corresponding metapeptide databases, and we downloaded the environmental non-redundant database (env_nr) from NCBI on Nov 12, 2022. We used Tide to build three peptide indices from the two metapeptide databases and from env_nr with default parameters, except we allowed three methionine oxidations per peptide and up to 1 missed cleavage. The resulting indices contained 41,665,963 peptides (CS), 34,116,884 peptides (BSt), and 310,021,565 peptides (env_nr), respectively, as well as

an equal number of shuffled decoy peptides. Each mzXML file was searched against the relevant metapeptide database and against env_nr using the tide-search command, specifying a precursor window of 10 ppm, allowing one isotope error, and using Tailor calibration. Finally, we used Percolator[29] with default parameters to jointly analyze the search results from each set of three runs against the same peptide index. All of the above commands were implemented within Crux[46] version 4.1-2fab3c91-2022-11-09. Metaproteomics spectrum files and peptides detected via database search are available on MassIVE with the dataset identifier MSV000094709 at https://doi.org/10.25345/C5ST7F78Z.

To estimate the error rate for the matching of Casanovo predicted peptides against a protein database, we developed a procedure akin to the target-decoy competition used in false discovery rate estimation for mass spectrometry database search[47]. To do so, we created a decoy protein database by randomly shuffling each protein sequence. We then asked whether each Casanovo prediction appears in the target (i.e., unshuffled) or decoy protein list, marking each predicted peptide as matching to the target, decoy, or neither. Peptides assigned to both were randomly assigned to either the target or decoy. We then segregated the peptides by length, and for each length we sorted the Casanovo predictions by confidence score. At each position k in the ranked list, we estimated the rate of random matching among the target matches as D_k/T_k, where D_k (respectively, T_k) is the number of decoys (respectively, targets) with rank smaller than k. We then selected the largest value of k such that $D_k/T_k < \alpha$, for some specified random matching rate α. In this work, we used $\alpha = 0.01$.

We used the online version of Prosit[32] at https://www.proteomicsdb.org/prosit/ to predict peak intensities and retention times using the Prosit_2020_intensity_hcd and Prosit_2019_irt models, respectively. A fixed collision energy of 27 was used for all peptides based on metadata from the spectrum files. Spectrum peaks predicted by Prosit were matched to observed peaks using a 0.05 Da fragment m/z tolerance to calculate the cosine similarity between the predicted and experimental spectra.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The revised nine-species benchmark is available on MassIVE with the dataset identifier MSV000090982 at https://doi.org/10.25345/C52V2CK8J. The non-enzymatic dataset with training, validation and test splits is available on MassIVE with the dataset identifier MSV000094014 at https://doi.org/10.25345/C5KS6JG0W. Metaproteomics spectrum files and peptides detected via database search are available on MassIVE with the dataset identifier MSV000094709 at https://doi.org/10.25345/C5ST7F78Z. The spectrum identifications for Casanovo and other evaluated tools on the revised nine-species benchmark are available on MassIVE with the dataset identifier MSV000094434 at https://doi.org/10.25345/C5B56DG2T. Similarly, Casanovo-only spectrum identifications for the immunopeptidomics, metaproteomics, and dark matter analyses are available on MassIVE with the dataset identifier MSV000093980 at https://doi.org/10.25345/C5VD6PG45. Source data are provided with this paper.

Code availability

Casanovo's source code and trained model weights from the MassIVE-KB training and non-enzymatic fine tuning are available under the Apache 2.0 license at https://github.com/Noble-Lab/casanovo[48]. Model weights from the nine-species benchmark training are available at https://doi.org/10.5281/zenodo.10694984. The transformer model was implemented using PyTorch[49] and PyTorch Lightning[50]. Additionally, NumPy[51], Pandas[52], and Scikit-Learn[53] were used for scientific data processing; and spectrum_utils[54], Pyteomics[49], and ppx[55] were used to process MS/MS data. Matplotlib[56] and Seaborn[57] were used for visualization purposes.

Casanovo's source code and trained model weights are available under the Apache 2.0 license at https://github.com/Noble-Lab/casanovo. Casanovo, Casanovo$_{bm}$ and Casanovo$_{ne}$ are all trained using version 4.0.1.

References

1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
2. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
3. Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A face in the crowd: recognizing peptides through database search. *Molecular and Cellular Proteomics* **10** (2011).
4. Bittremieux, W. et al. Quality control in mass spectrometry-based proteomics. *Mass Spectrom. Rev.* **37**, 697–711 (2018).
5. Stopfer, L. E., Mesfin, J. M., Joughin, B. A., Lauffenburger, D. A. & White, F. M. Multiplexed relative and absolute quantitative immunopeptidomics reveals MHC I repertoire alterations induced by CDK4/6 inhibition. *Nat. Commun.* **11**, 1–14 (2020).
6. Mayer, R. L. & Impens, F. Immunopeptidomics for next-generation bacterial vaccine development. *Trends Microbiol.* **29**, 1034–1045 (2021).
7. Hunt, D. F. et al. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Sci. (N. Y., N. Y.)* **255**, 1261–1263 (1992).
8. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. USA* **31**, 8247–8252 (2017).
9. Muth, T., Benndorf, D., Reichl, U., Rapp, E. & Martens, L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.* **9**, 578–585 (2013).
10. Taylor, J. A. & Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**, 1067–1075 (1997).
11. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
12. Dancik, V., Addona, T., Clauser, K., Vath, J. & Pevzner, P. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6**, 327–342 (1999).
13. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
14. Fischer, B. et al. A hidden Markov model for de novo peptide sequencing. *Adv. Neural Inf. Process. Syst.* **17**, 457–464 (2005).
15. Ma, B. Novor: Real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885–1894 (2015).
16. Karunratanakul, K., Tang, H.-Y., Speicher, D. W., Chuangsuwanich, E. & Sriswasdi, S. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Mol. Cell. Proteom.* **18**, 2478–2491 (2019).
17. Qiao, R. et al. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat. Mach. Intell.* **3**, 420–425 (2021).
18. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. PointNet: deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, 652–660 (2016).

19. Yang, H., Chi, H., Zeng, W., Zhou, W. & He, S. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* **35**, i83–i90 (2019).
20. Zhou, X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
21. Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems***30** (2017).
22. Yilmaz, M., Fondrie, W. E., Bittremieux, W., Oh, S. & Noble, W. S. *De novo* mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the International Conference on Machine Learning*, 25514–25522 (2022).
23. Wang, M. et al. Assembling the community-scale discoverable human proteome. *Cell Syst.* **7**, 412–421.e5 (2018).
24. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).
25. Avsec, Z. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
26. Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *Nat. Methods* **19**, 675–678 (2022).
27. Shouman, O., Gabriel, W., Giurcoiu, V.-G., Sternlicht, V. & Wilhelm, M. Prospect: Labeled tandem mass spectrometry dataset for machine learning in proteomics. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).
28. Diament, B. & Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **10**, 3871–3879 (2011).
29. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
30. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
31. May, D. H. et al. An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteome Res.* **15**, 2697–2705 (2016).
32. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509 (2019).
33. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
34. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186 (2019).
35. Johnson, R. S. et al. Assessing protein sequence database suitability using de novo sequencing. *Mol. Cell. Proteom.* **19**, 198–208 (2020).
36. Kertesz-Farkas, A., Keich, U. & Noble, W. S. Tandem mass spectrum identification via cascaded search. *J. Proteome Res.* **14**, 3027–3038 (2015).
37. Beslic, D., Tscheuschner, G., Renard, B. Y., Weller, M. G. & Muth, T. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Briefings in Bioinoformatics* (2022). Advance online access.
38. Mao, Z., Zhang, R., Xin, L. & Li, M. Mitigating the missing fragmentation problem in de novo peptide sequencing with a two stage graph-based deep learning model. *Nature Machine Intelligence* **5** (2023).
39. Liu, K., Ye, Y., Li, S. & Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nat. Commun.* **14**, 7974 (2023).
40. Wu, R., Zhang, X., Wang, R. & Wang, H. Denovo-GCN: De novo peptide sequencing by graph convolutional neural networks. *Applied Sciences* **13** (2023).
41. Klaproth-Andrade, D. et al. Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. *Nat. Commun.* **15**, 151 (2024).
42. Yang, T. et al. Introducing π-HelixNovo for practical large-scale de novo peptide sequencing. *Brief. Bioinforma.* **25**, bbae021 (2024).
43. Lee, S. & Kim, H. Bidirectional de novo peptide sequencing using a transformer model. *PLOS Computational Biol.* **20**, e1011892 (2024).
44. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).
45. Sulimov, P. & Kertész-Farkas, A. Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. *J. Proteome Res.* **19**, 1481–1490 (2020).
46. Park, C. Y., Klammer, A. A., Käll, L., MacCoss, M. P. & Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7**, 3022–3027 (2008).
47. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
48. Yilmaz, M. Noble-lab/casanovo (2023). https://doi.org/10.5281/zenodo.11205039.
49. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., Vancouver, Canada, 2019).
50. Falcon, W. & Team, T. PyTorch Lightning the lightweight PyTorch wrapper for high-performance AI research. scale your models, not the boilerplate (2019).
51. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
52. McKinney, W. Data structures for statistical computing in Python. in Proceedings of the 9th Python in Science Conference (eds. van der Walt, S. & Millman, J.) 51–56 (Austin, Texas, USA, 2020).
53. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
54. Bittremieux, W. spectrum_utils: A python package for mass spectrometry data processing and visualization. *Anal. Chem.* **92**, 659–661 (2020).
55. Fondrie, W., Bittremieux, W. & Noble, W. S. ppx: Programmatic access to proteomics data repositories. *J. Proteome Res.* **20**, 4621–4624 (2021).
56. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
57. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements

This work is in part supported by CCF-2019844 (S.O.) as a part of the NSF Institute for Foundations of Machine Learning (IFML) as well as by National Science Foundation award 2245300 (W.S.N.).

Author contributions

W.S.N., M.Y., W.B., and W.E.F. conceptualized the work. W.S.N. and S.O. supervised the work. M.Y., W.B., and W.E.F. developed the Casanovo software. M.Y., C.F.M., V.A., W.B., and R.N. carried out the analyses. W.S.N., M.Y., W.B., and W.E.F. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests
The authors declare no conflicts of interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-49731-x.

Correspondence and requests for materials should be addressed to William Stafford Noble.

Peer review information *Nature Communications* thanks Ekapol Chuangsuwanich, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024, modified publication 2026

A.9 Protein structure

Recommended reading for Chapter 9 (listed in Section 9.14).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Anfinsen (1973) — why the sequence determines the structure; the basis for all structure prediction.
  DOI · via SU library (SU login)

  Principles that Govern the Folding of Protein Chains. Anfinsen (1973). *Science*. doi:10.1126/science.181.4096.223

- Marks et al. (2012) — protein structure prediction from sequence variation; coevolution explained.
  DOI · via SU library (SU login)

  Protein structure prediction from sequence variation. Marks et al. (2012). *Nature Biotechnology*. doi:10.1038/nbt.2419

  > Genomic sequences contain rich evolutionary information about functional constraints on macromolecules such as proteins. This information can be efficiently mined to detect evolutionary couplings between residues in proteins and address the long-standing challenge to compute protein three-dimensional structures from amino acid sequences. Substantial progress has recently been made on this problem owing to the explosive growth in available sequences and the application of global statistical methods. In addition to three-dimensional structure, the improved understanding of covariation may help identify functional residues involved in ligand binding, protein-complex formation and conformational changes. We expect computation of covariation patterns to complement experimental structural biology in elucidating the full spectrum of protein structures, their functional interactions and evolutionary dynamics.

- Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. (2021) — AlphaFold2; the method that largely solved single-chain structure prediction.
  full text below · DOI · open access (CC BY 4.0)

  Highly accurate protein structure prediction with AlphaFold. Jumper et al. (2021). *Nature*. doi:10.1038/s41586-021-03819-2

  > Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort 1-4 , the structures of around 100,000 unique proteins have been determined 5 , but this represents a small fraction of the billions of known protein sequences 6,7 . Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence-the structure prediction component of the 'protein folding problem' 8 -has been an important open research problem for more than 50 years 9 . Despite recent progress 10-14 , existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14) 15 , demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

- Abramson et al. (2024) — AlphaFold3; co-folding of proteins with nucleic acids, ligands and ions by diffusion.
  full text below · DOI · open access (CC BY 4.0)

Accurate structure prediction of biomolecular interactions with AlphaFold 3. Abramson et al. (2024). *Nature*. doi:10.1038/s41586-024-07487-w

> The introduction of AlphaFold 2 1 has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design 2-6 . Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy over many previous specialized tools: far greater accuracy for protein-ligand interactions compared with state-of-the-art docking tools, much higher accuracy for protein-nucleic acid interactions compared with nucleic-acid-specific predictors and substantially higher antibody-antigen prediction accuracy compared with AlphaFold-Multimer v.2.3 7,8 . Together, these results show that high-accuracy modelling across biomolecular space is possible within a single unified deep-learning framework.

- Lin et al. (2023) — ESMFold and the ESM Metagenomic Atlas; structure from a language model.
  DOI · via SU library (SU login)

  Evolutionary-scale prediction of atomic-level protein structure with a language model. Lin et al. (2023). *Science*. doi:10.1126/science.ade2574

  > Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

- Watson et al. (2023) — RFdiffusion; de novo design of protein structure and function with diffusion models.
  full text below · DOI · open access (CC BY 4.0)

  De novo design of protein structure and function with RFdiffusion. Watson et al. (2023). *Nature*. doi:10.1038/s41586-023-06415-8

  > There has been considerable recent progress in designing new proteins using deep-learning methods 1-9 . Despite this progress, a general deep-learning framework for protein design that enables solution of a wide range of design challenges, including de novo binder design and design of higher-order symmetric architectures, has yet to be described. Diffusion models 10,11 have had considerable success in image and language generative modelling but limited success when applied to protein modelling, probably due to the complexity of protein backbone geometry and sequence-structure relationships. Here we show that by fine-tuning the RoseTTAFold structure prediction network on protein structure denoising tasks, we obtain a generative model of protein backbones that achieves outstanding performance on unconditional and topology-constrained protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding and symmetric motif scaffolding for therapeutic and metal-binding protein design. We demonstrate the power and generality of the method, called RoseTTAFold diffusion (RFdiffusion), by experimentally characterizing the structures and functions of hundreds of designed symmetric assemblies, metal-binding proteins and protein binders. The accuracy of RFdiffusion is confirmed by the cryogenic electron microscopy structure of a designed binder in complex with influenza haemagglutinin that is nearly identical to the design model. In a manner analogous to networks that produce images from user-specified inputs, RFdiffusion enables the design of diverse functional proteins from simple molecular specifications.

The open-access papers are reproduced below under their original licences.

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

Open access

Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

The development of computational methods to predict three-dimensional (3D) protein structures from the protein sequence has proceeded along two complementary paths that focus on either the physical interactions or the evolutionary history. The physical interaction programme heavily integrates our understanding of molecular driving forces into either thermodynamic or kinetic simulation of protein physics[16] or statistical approximations thereof[17]. Although theoretically very appealing, this approach has proved highly challenging for even moderate-sized proteins due to the computational intractability of molecular simulation, the context dependence of protein stability and the difficulty of producing sufficiently accurate models of protein physics. The evolutionary programme has provided an alternative in recent years, in which the constraints on protein structure are derived from bioinformatics analysis of the evolutionary history of proteins, homology to solved structures[18,19] and pairwise evolutionary correlations[20–24]. This bioinformatics approach has benefited greatly from the steady growth of experimental protein structures deposited in the Protein Data Bank (PDB)[5], the explosion of genomic sequencing and the rapid development of deep learning techniques to interpret these correlations. Despite these advances, contemporary physical and evolutionary-history-based approaches produce predictions that are far short of experimental accuracy in the majority of cases in which a close homologue has not been solved experimentally and this has limited their utility for many biological applications.

In this study, we develop the first, to our knowledge, computational approach capable of predicting protein structures to near experimental accuracy in a majority of cases. The neural network AlphaFold that we developed was entered into the CASP14 assessment (May–July 2020; entered under the team name 'AlphaFold2' and a completely different model from our CASP13 AlphaFold system[10]). The CASP assessment is carried out biennially using recently solved structures that have not been deposited in the PDB or publicly disclosed so that it is a blind test

[1]DeepMind, London, UK. [2]School of Biological Sciences, Seoul National University, Seoul, South Korea. [3]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea. [4]These authors contributed equally: John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Demis Hassabis.
✉e-mail: jumper@deepmind.com; dhcontact@deepmind.com

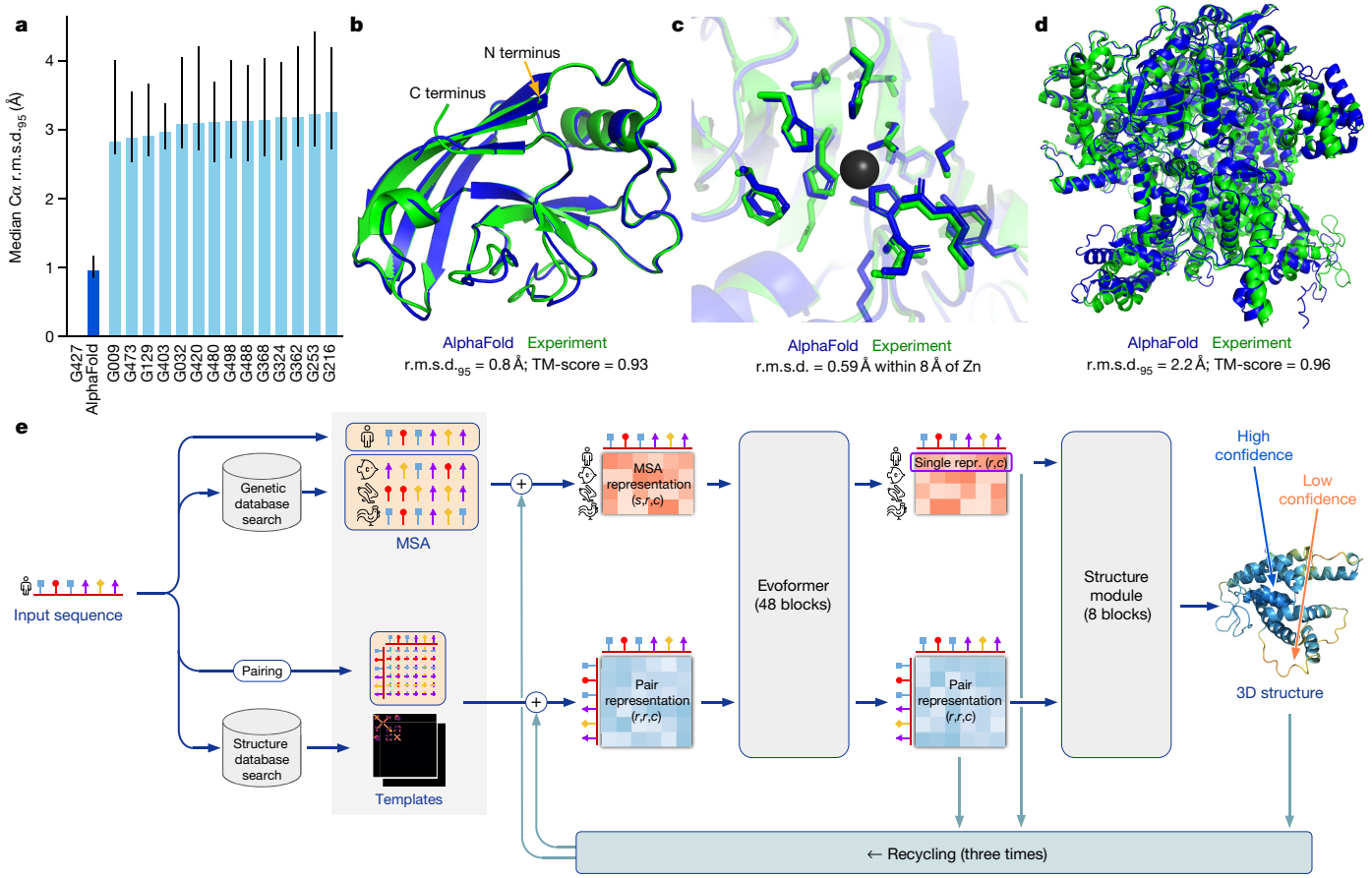


Fig. 1 | AlphaFold produces highly accurate structures. a, The performance of AlphaFold on the CASP14 dataset ($n = 87$ protein domains) relative to the top-15 entries (out of 146 entries), group numbers correspond to the numbers assigned to entrants by CASP. Data are median and the 95% confidence interval of the median, estimated from 10,000 bootstrap samples. **b**, Our prediction of CASP14 target T1049 (PDB 6Y4F, blue) compared with the true (experimental) structure (green). Four residues in the C terminus of the crystal structure are *B*-factor outliers and are not depicted. **c**, CASP14 target T1056 (PDB 6YJ1). An example of a well-predicted zinc-binding site (AlphaFold has accurate side chains even though it does not explicitly predict the zinc ion). **d**, CASP target T1044 (PDB 6VR4)—a 2,180-residue single chain—was predicted with correct domain packing (the prediction was made after CASP using AlphaFold without intervention). **e**, Model architecture. Arrows show the information flow among the various components described in this paper. Array shapes are shown in parentheses with *s*, number of sequences (N_{seq} in the main text); *r*, number of residues (N_{res} in the main text); *c*, number of channels.

for the participating methods, and has long served as the gold-standard assessment for the accuracy of structure prediction[25,26].

In CASP14, AlphaFold structures were vastly more accurate than competing methods. AlphaFold structures had a median backbone accuracy of 0.96 Å r.m.s.d.$_{95}$ (Cα root-mean-square deviation at 95% residue coverage) (95% confidence interval = 0.85–1.16 Å) whereas the next best performing method had a median backbone accuracy of 2.8 Å r.m.s.d.$_{95}$ (95% confidence interval = 2.7–4.0 Å) (measured on CASP domains; see Fig. 1a for backbone accuracy and Supplementary Fig. 14 for all-atom accuracy). As a comparison point for this accuracy, the width of a carbon atom is approximately 1.4 Å. In addition to very accurate domain structures (Fig. 1b), AlphaFold is able to produce highly accurate side chains (Fig. 1c) when the backbone is highly accurate and considerably improves over template-based methods even when strong templates are available. The all-atom accuracy of AlphaFold was 1.5 Å r.m.s.d.$_{95}$ (95% confidence interval = 1.2–1.6 Å) compared with the 3.5 Å r.m.s.d.$_{95}$ (95% confidence interval = 3.1–4.2 Å) of the best alternative method. Our methods are scalable to very long proteins with accurate domains and domain-packing (see Fig. 1d for the prediction of a 2,180-residue protein with no structural homologues). Finally, the model is able to provide precise, per-residue estimates of its reliability that should enable the confident use of these predictions.

We demonstrate in Fig. 2a that the high accuracy that AlphaFold demonstrated in CASP14 extends to a large sample of recently released PDB structures; in this dataset, all structures were deposited in the PDB after our training data cut-off and are analysed as full chains (see Methods, Supplementary Fig. 15 and Supplementary Table 6 for more details). Furthermore, we observe high side-chain accuracy when the backbone prediction is accurate (Fig. 2b) and we show that our confidence measure, the predicted local-distance difference test (pLDDT), reliably predicts the Cα local-distance difference test (lDDT-Cα) accuracy of the corresponding prediction (Fig. 2c). We also find that the global superposition metric template modelling score (TM-score)[27] can be accurately estimated (Fig. 2d). Overall, these analyses validate that the high accuracy and reliability of AlphaFold on CASP14 proteins also transfers to an uncurated collection of recent PDB submissions, as would be expected (see Supplementary Methods 1.15 and Supplementary Fig. 11 for confirmation that this high accuracy extends to new folds).

The AlphaFold network

AlphaFold greatly improves the accuracy of structure prediction by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures. In particular, we demonstrate a new architecture to jointly embed multiple sequence alignments (MSAs) and pairwise features, a new output representation and associated loss that enable accurate end-to-end structure prediction, a new equivariant attention

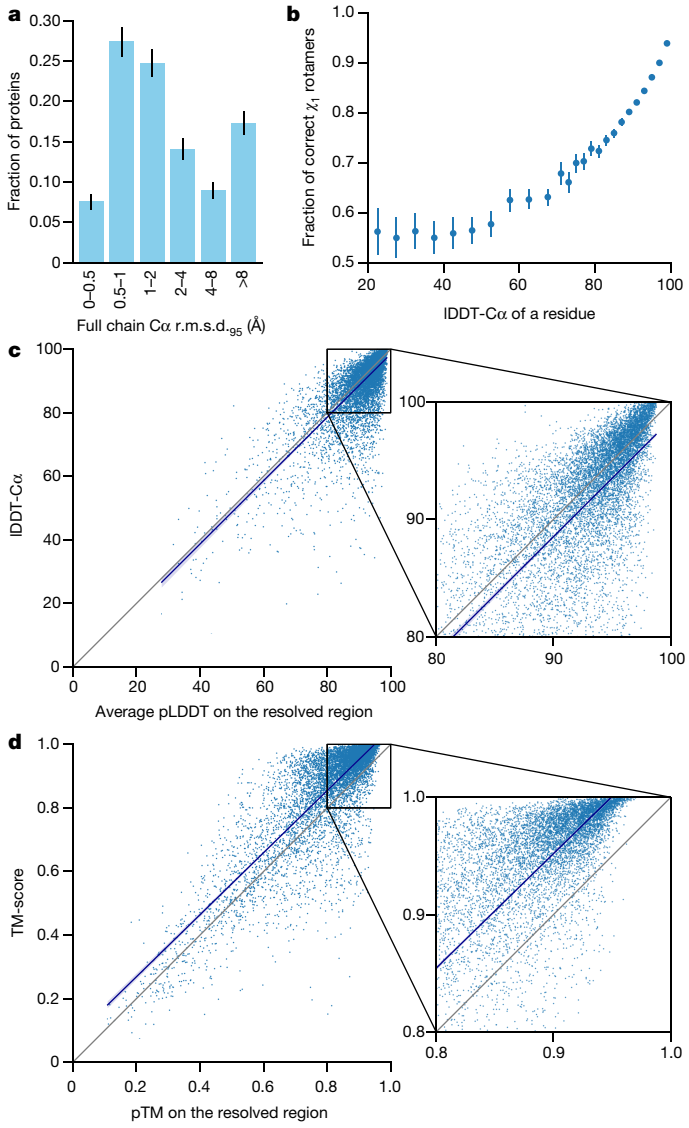


Fig. 2 | Accuracy of AlphaFold on recent PDB structures. The analysed structures are newer than any structure in the training set. Further filtering is applied to reduce redundancy (see Methods). **a**, Histogram of backbone r.m.s.d. for full chains (Cα r.m.s.d. at 95% coverage). Error bars are 95% confidence intervals (Poisson). This dataset excludes proteins with a template (identified by hmmsearch) from the training set with more than 40% sequence identity covering more than 1% of the chain (n = 3,144 protein chains). The overall median is 1.46 Å (95% confidence interval = 1.40–1.56 Å). Note that this measure will be highly sensitive to domain packing and domain accuracy; a high r.m.s.d. is expected for some chains with uncertain packing or packing errors. **b**, Correlation between backbone accuracy and side-chain accuracy. Filtered to structures with any observed side chains and resolution better than 2.5 Å (n = 5,317 protein chains); side chains were further filtered to B-factor <30 Å^2. A rotamer is classified as correct if the predicted torsion angle is within 40°. Each point aggregates a range of lDDT-Cα, with a bin size of 2 units above 70 lDDT-Cα and 5 units otherwise. Points correspond to the mean accuracy; error bars are 95% confidence intervals (Student t-test) of the mean on a per-residue basis. **c**, Confidence score compared to the true accuracy on chains. Least-squares linear fit lDDT-Cα = 0.997 × pLDDT − 1.17 (Pearson's r = 0.76). n = 10,795 protein chains. The shaded region of the linear fit represents a 95% confidence interval estimated from 10,000 bootstrap samples. In the companion paper[39], additional quantification of the reliability of pLDDT as a confidence measure is provided. **d**, Correlation between pTM and full chain TM-score. Least-squares linear fit TM-score = 0.98 × pTM + 0.07 (Pearson's r = 0.85). n = 10,795 protein chains. The shaded region of the linear fit represents a 95% confidence interval estimated from 10,000 bootstrap samples.

architecture, use of intermediate losses to achieve iterative refinement of predictions, masked MSA loss to jointly train with the structure, learning from unlabelled protein sequences using self-distillation and self-estimates of accuracy.

The AlphaFold network directly predicts the 3D coordinates of all heavy atoms for a given protein using the primary amino acid sequence and aligned sequences of homologues as inputs (Fig. 1e; see Methods for details of inputs including databases, MSA construction and use of templates). A description of the most important ideas and components is provided below. The full network architecture and training procedure are provided in the Supplementary Methods.

The network comprises two main stages. First, the trunk of the network processes the inputs through repeated layers of a novel neural network block that we term Evoformer to produce an $N_{seq} \times N_{res}$ array (N_{seq}, number of sequences; N_{res}, number of residues) that represents a processed MSA and an $N_{res} \times N_{res}$ array that represents residue pairs. The MSA representation is initialized with the raw MSA (although see Supplementary Methods 1.2.7 for details of handling very deep MSAs). The Evoformer blocks contain a number of attention-based and non-attention-based components. We show evidence in 'Interpreting the neural network' that a concrete structural hypothesis arises early within the Evoformer blocks and is continuously refined. The key innovations in the Evoformer block are new mechanisms to exchange information within the MSA and pair representations that enable direct reasoning about the spatial and evolutionary relationships.

The trunk of the network is followed by the structure module that introduces an explicit 3D structure in the form of a rotation and translation for each residue of the protein (global rigid body frames). These representations are initialized in a trivial state with all rotations set to the identity and all positions set to the origin, but rapidly develop and refine a highly accurate protein structure with precise atomic details. Key innovations in this section of the network include breaking the chain structure to allow simultaneous local refinement of all parts of the structure, a novel equivariant transformer to allow the network to implicitly reason about the unrepresented side-chain atoms and a loss term that places substantial weight on the orientational correctness of the residues. Both within the structure module and throughout the whole network, we reinforce the notion of iterative refinement by repeatedly applying the final loss to outputs and then feeding the outputs recursively into the same modules. The iterative refinement using the whole network (which we term 'recycling' and is related to approaches in computer vision[28,29]) contributes markedly to accuracy with minor extra training time (see Supplementary Methods 1.8 for details).

Evoformer

The key principle of the building block of the network—named Evoformer (Figs. 1e, 3a)—is to view the prediction of protein structures as a graph inference problem in 3D space in which the edges of the graph are defined by residues in proximity. The elements of the pair representation encode information about the relation between the residues (Fig. 3b). The columns of the MSA representation encode the individual residues of the input sequence while the rows represent the sequences in which those residues appear. Within this framework, we define a number of update operations that are applied in each block in which the different update operations are applied in series.

The MSA representation updates the pair representation through an element-wise outer product that is summed over the MSA sequence dimension. In contrast to previous work[30], this operation is applied within every block rather than once in the network, which enables the continuous communication from the evolving MSA representation to the pair representation.

Within the pair representation, there are two different update patterns. Both are inspired by the necessity of consistency of the pair

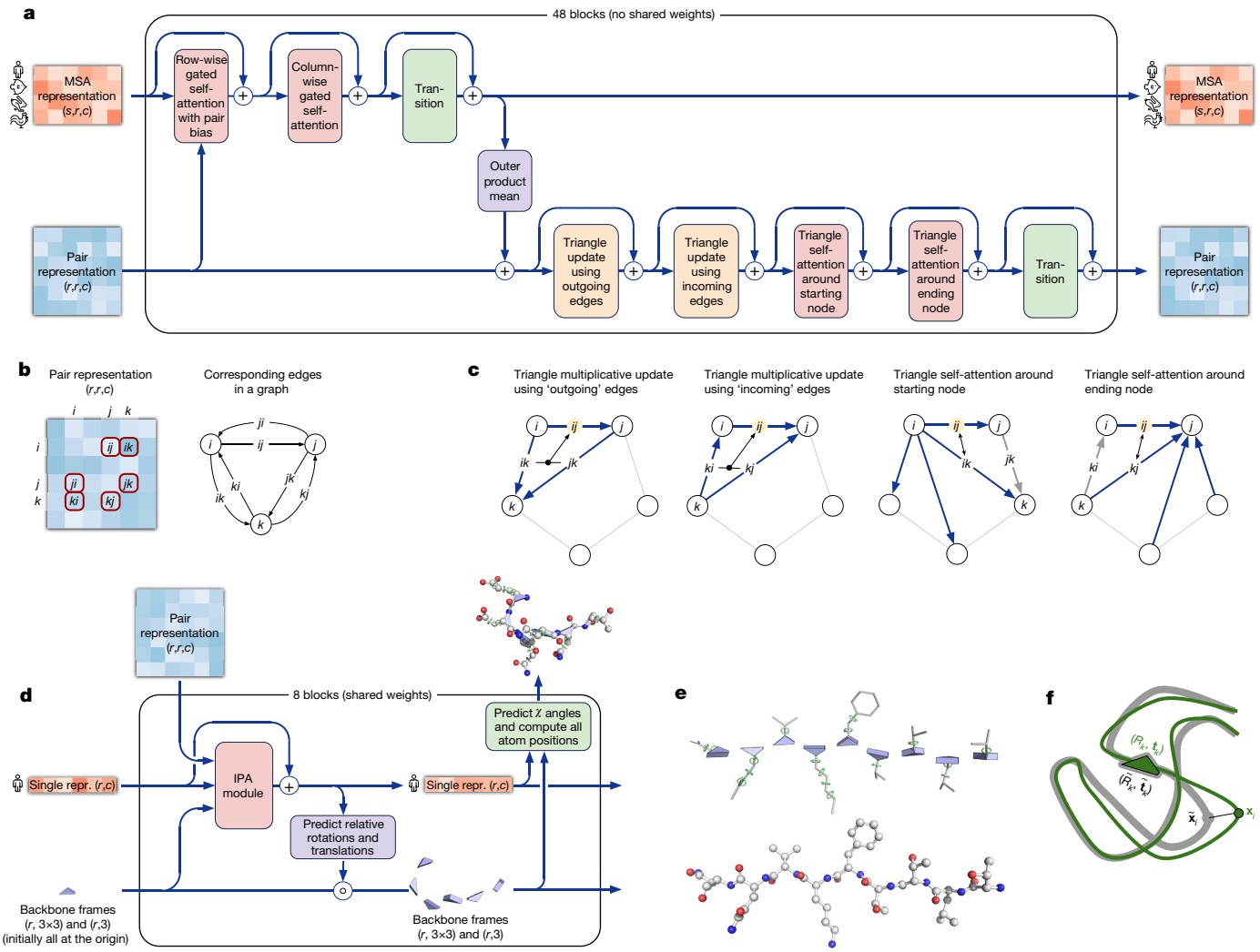


Fig. 3 | Architectural details. a, Evoformer block. Arrows show the information flow. The shape of the arrays is shown in parentheses. **b**, The pair representation interpreted as directed edges in a graph. **c**, Triangle multiplicative update and triangle self-attention. The circles represent residues. Entries in the pair representation are illustrated as directed edges and in each diagram, the edge being updated is *ij*. **d**, Structure module including Invariant point attention (IPA) module. The single representation is a copy of the first row of the MSA representation. **e**, Residue gas: a representation of each residue as one free-floating rigid body for the backbone (blue triangles) and χ angles for the side chains (green circles). The corresponding atomic structure is shown below. **f**, Frame aligned point error (FAPE). Green, predicted structure; grey, true structure; $(R_k, \mathbf{t}_k)$, frames; $\mathbf{x}_i$, atom positions.

representation—for a pairwise description of amino acids to be representable as a single 3D structure, many constraints must be satisfied including the triangle inequality on distances. On the basis of this intuition, we arrange the update operations on the pair representation in terms of triangles of edges involving three different nodes (Fig. 3c). In particular, we add an extra logit bias to axial attention[31] to include the 'missing edge' of the triangle and we define a non-attention update operation 'triangle multiplicative update' that uses two edges to update the missing third edge (see Supplementary Methods 1.6.5 for details). The triangle multiplicative update was developed originally as a more symmetric and cheaper replacement for the attention, and networks that use only the attention or multiplicative update are both able to produce high-accuracy structures. However, the combination of the two updates is more accurate.

We also use a variant of axial attention within the MSA representation. During the per-sequence attention in the MSA, we project additional logits from the pair stack to bias the MSA attention. This closes the loop by providing information flow from the pair representation back into the MSA representation, ensuring that the overall Evoformer block is able to fully mix information between the pair and MSA representations and prepare for structure generation within the structure module.

End-to-end structure prediction

The structure module (Fig. 3d) operates on a concrete 3D backbone structure using the pair representation and the original sequence row (single representation) of the MSA representation from the trunk. The 3D backbone structure is represented as N_{res} independent rotations and translations, each with respect to the global frame (residue gas) (Fig. 3e). These rotations and translations—representing the geometry of the N-Cα-C atoms—prioritize the orientation of the protein backbone so that the location of the side chain of each residue is highly constrained within that frame. Conversely, the peptide bond geometry is completely unconstrained and the network is observed to frequently violate the chain constraint during the application of the structure module as breaking this constraint enables the local refinement of all parts of the chain without solving complex loop closure problems. Satisfaction of the peptide bond geometry is encouraged during fine-tuning by a violation loss term. Exact enforcement of peptide bond geometry is only achieved in the post-prediction relaxation of the structure by gradient descent in the Amber[32] force field. Empirically, this final relaxation does not improve the accuracy of the model as measured by the

global distance test (GDT)[33] or lDDT-Cα[34] but does remove distracting stereochemical violations without the loss of accuracy.

The residue gas representation is updated iteratively in two stages (Fig. 3d). First, a geometry-aware attention operation that we term 'invariant point attention' (IPA) is used to update an N_{res} set of neural activations (single representation) without changing the 3D positions, then an equivariant update operation is performed on the residue gas using the updated activations. The IPA augments each of the usual attention queries, keys and values with 3D points that are produced in the local frame of each residue such that the final value is invariant to global rotations and translations (see Methods 'IPA' for details). The 3D queries and keys also impose a strong spatial/locality bias on the attention, which is well-suited to the iterative refinement of the protein structure. After each attention operation and element-wise transition block, the module computes an update to the rotation and translation of each backbone frame. The application of these updates within the local frame of each residue makes the overall attention and update block an equivariant operation on the residue gas.

Predictions of side-chain χ angles as well as the final, per-residue accuracy of the structure (pLDDT) are computed with small per-residue networks on the final activations at the end of the network. The estimate of the TM-score (pTM) is obtained from a pairwise error prediction that is computed as a linear projection from the final pair representation. The final loss (which we term the frame-aligned point error (FAPE) (Fig. 3f)) compares the predicted atom positions to the true positions under many different alignments. For each alignment, defined by aligning the predicted frame $(R_k, \mathbf{t}_k)$ to the corresponding true frame, we compute the distance of all predicted atom positions $\mathbf{x}_i$ from the true atom positions. The resulting $N_{frames} \times N_{atoms}$ distances are penalized with a clamped L^1 loss. This creates a strong bias for atoms to be correct relative to the local frame of each residue and hence correct with respect to its side-chain interactions, as well as providing the main source of chirality for AlphaFold (Supplementary Methods 1.9.3 and Supplementary Fig. 9).

Training with labelled and unlabelled data

The AlphaFold architecture is able to train to high accuracy using only supervised learning on PDB data, but we are able to enhance accuracy (Fig. 4a) using an approach similar to noisy student self-distillation[35]. In this procedure, we use a trained network to predict the structure of around 350,000 diverse sequences from Uniclust30[36] and make a new dataset of predicted structures filtered to a high-confidence subset. We then train the same architecture again from scratch using a mixture of PDB data and this new dataset of predicted structures as the training data, in which the various training data augmentations such as cropping and MSA subsampling make it challenging for the network to recapitulate the previously predicted structures. This self-distillation procedure makes effective use of the unlabelled sequence data and considerably improves the accuracy of the resulting network.

Additionally, we randomly mask out or mutate individual residues within the MSA and have a Bidirectional Encoder Representations from Transformers (BERT)-style[37] objective to predict the masked elements of the MSA sequences. This objective encourages the network to learn to interpret phylogenetic and covariation relationships without hardcoding a particular correlation statistic into the features. The BERT objective is trained jointly with the normal PDB structure loss on the same training examples and is not pre-trained, in contrast to recent independent work[38].

Interpreting the neural network

To understand how AlphaFold predicts protein structure, we trained a separate structure module for each of the 48 Evoformer blocks in the network while keeping all parameters of the main network frozen (Supplementary Methods 1.14). Including our recycling stages, this provides a trajectory of 192 intermediate structures—one per full

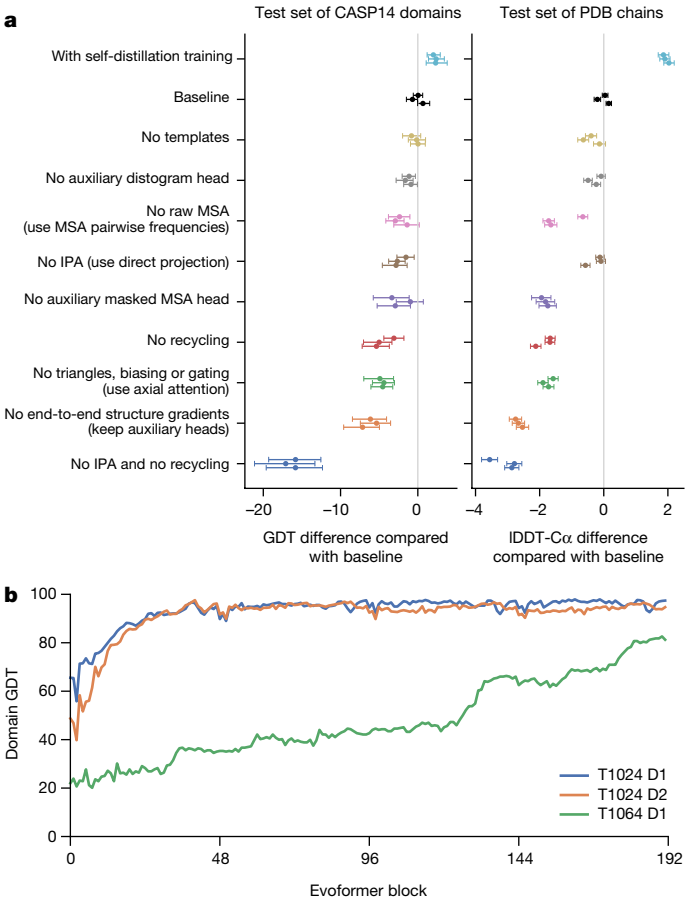


Fig. 4 | Interpreting the neural network. a, Ablation results on two target sets: the CASP14 set of domains ($n = 87$ protein domains) and the PDB test set of chains with template coverage of ≤30% at 30% identity ($n = 2,261$ protein chains). Domains are scored with GDT and chains are scored with lDDT-Cα. The ablations are reported as a difference compared with the average of the three baseline seeds. Means (points) and 95% bootstrap percentile intervals (error bars) are computed using bootstrap estimates of 10,000 samples. **b**, Domain GDT trajectory over 4 recycling iterations and 48 Evoformer blocks on CASP14 targets LmrP (T1024) and Orf8 (T1064) where D1 and D2 refer to the individual domains as defined by the CASP assessment. Both T1024 domains obtain the correct structure early in the network, whereas the structure of T1064 changes multiple times and requires nearly the full depth of the network to reach the final structure. Note, 48 Evoformer blocks comprise one recycling iteration.

Evoformer block—in which each intermediate represents the belief of the network of the most likely structure at that block. The resulting trajectories are surprisingly smooth after the first few blocks, showing that AlphaFold makes constant incremental improvements to the structure until it can no longer improve (see Fig. 4b for a trajectory of accuracy). These trajectories also illustrate the role of network depth. For very challenging proteins such as ORF8 of SARS-CoV-2 (T1064), the network searches and rearranges secondary structure elements for many layers before settling on a good structure. For other proteins such as LmrP (T1024), the network finds the final structure within the first few layers. Structure trajectories of CASP14 targets T1024, T1044, T1064 and T1091 that demonstrate a clear iterative building process for a range of protein sizes and difficulties are shown in Supplementary Videos 1–4. In Supplementary Methods 1.16 and Supplementary Figs. 12, 13, we interpret the attention maps produced by AlphaFold layers.

Figure 4a contains detailed ablations of the components of AlphaFold that demonstrate that a variety of different mechanisms contribute to AlphaFold accuracy. Detailed descriptions of each ablation model, their training details, extended discussion of ablation results and the

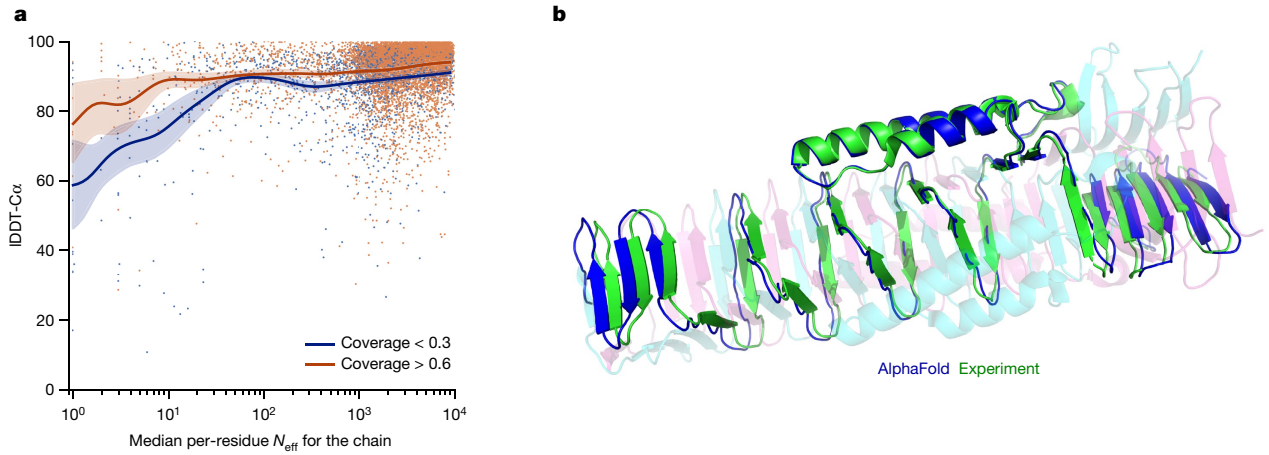


Fig. 5 | Effect of MSA depth and cross-chain contacts. a, Backbone accuracy (lDDT-Cα) for the redundancy-reduced set of the PDB after our training data cut-off, restricting to proteins in which at most 25% of the long-range contacts are between different heteromer chains. We further consider two groups of proteins based on template coverage at 30% sequence identity: covering more than 60% of the chain (n = 6,743 protein chains) and covering less than 30% of the chain (n = 1,596 protein chains). MSA depth is computed by counting the number of non-gap residues for each position in the MSA (using the N_{eff} weighting scheme; see Methods for details) and taking the median across residues. The curves are obtained through Gaussian kernel average smoothing (window size is 0.2 units in $\log_{10}(N_{eff})$); the shaded area is the 95% confidence interval estimated using bootstrap of 10,000 samples. **b**, An intertwined homotrimer (PDB 6SK0) is correctly predicted without input stoichiometry and only a weak template (blue is predicted and green is experimental).

effect of MSA depth on each ablation are provided in Supplementary Methods 1.13 and Supplementary Fig. 10.

MSA depth and cross-chain contacts

Although AlphaFold has a high accuracy across the vast majority of deposited PDB structures, we note that there are still factors that affect accuracy or limit the applicability of the model. The model uses MSAs and the accuracy decreases substantially when the median alignment depth is less than around 30 sequences (see Fig. 5a for details). We observe a threshold effect where improvements in MSA depth over around 100 sequences lead to small gains. We hypothesize that the MSA information is needed to coarsely find the correct structure within the early stages of the network, but refinement of that prediction into a high-accuracy model does not depend crucially on the MSA information. The other substantial limitation that we have observed is that AlphaFold is much weaker for proteins that have few intra-chain or homotypic contacts compared to the number of heterotypic contacts (further details are provided in a companion paper[39]). This typically occurs for bridging domains within larger complexes in which the shape of the protein is created almost entirely by interactions with other chains in the complex. Conversely, AlphaFold is often able to give high-accuracy predictions for homomers, even when the chains are substantially intertwined (Fig. 5b). We expect that the ideas of AlphaFold are readily applicable to predicting full hetero-complexes in a future system and that this will remove the difficulty with protein chains that have a large number of hetero-contacts.

Related work

The prediction of protein structures has had a long and varied development, which is extensively covered in a number of reviews[14,40–43]. Despite the long history of applying neural networks to structure prediction[14,42,43], they have only recently come to improve structure prediction[10,11,44,45]. These approaches effectively leverage the rapid improvement in computer vision systems[46] by treating the problem of protein structure prediction as converting an 'image' of evolutionary couplings[22–24] to an 'image' of the protein distance matrix and then integrating the distance predictions into a heuristic system that produces the final 3D coordinate prediction. A few recent studies have been developed to predict the 3D coordinates directly[47–50], but the accuracy of these approaches does not match traditional, hand-crafted structure prediction pipelines[51]. In parallel, the success of attention-based networks for language processing[52] and, more recently, computer vision[31,53] has inspired the exploration of attention-based methods for interpreting protein sequences[54–56].

Discussion

The methodology that we have taken in designing AlphaFold is a combination of the bioinformatics and physical approaches: we use a physical and geometric inductive bias to build components that learn from PDB data with minimal imposition of handcrafted features (for example, AlphaFold builds hydrogen bonds effectively without a hydrogen bond score function). This results in a network that learns far more efficiently from the limited data in the PDB but is able to cope with the complexity and variety of structural data.

In particular, AlphaFold is able to handle missing the physical context and produce accurate models in challenging cases such as intertwined homomers or proteins that only fold in the presence of an unknown haem group. The ability to handle underspecified structural conditions is essential to learning from PDB structures as the PDB represents the full range of conditions in which structures have been solved. In general, AlphaFold is trained to produce the protein structure most likely to appear as part of a PDB structure. For example, in cases in which a particular stochiometry, ligand or ion is predictable from the sequence alone, AlphaFold is likely to produce a structure that respects those constraints implicitly.

AlphaFold has already demonstrated its utility to the experimental community, both for molecular replacement[57] and for interpreting cryogenic electron microscopy maps[58]. Moreover, because AlphaFold outputs protein coordinates directly, AlphaFold produces predictions in graphics processing unit (GPU) minutes to GPU hours depending on the length of the protein sequence (for example, around one GPU minute per model for 384 residues; see Methods for details). This opens up the exciting possibility of predicting structures at the proteome-scale and beyond—in a companion paper[39], we demonstrate the application of AlphaFold to the entire human proteome[39].

The explosion in available genomic sequencing techniques and data has revolutionized bioinformatics but the intrinsic challenge of experimental structure determination has prevented a similar expansion in our structural knowledge. By developing an accurate protein structure

prediction algorithm, coupled with existing large and well-curated structure and sequence databases assembled by the experimental community, we hope to accelerate the advancement of structural bioinformatics that can keep pace with the genomics revolution. We hope that AlphaFold—and computational approaches that apply its techniques for other biophysical problems—will become essential tools of modern biology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03819-2.

1. Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomic resolution macromolecular structure determination. *F1000Res*. **9**, 667 (2020).
2. Bai, X.-C., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci*. **40**, 49–57 (2015).
3. Jaskolski, M., Dauter, Z. & Wlodawer, A. A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *FEBS J*. **281**, 3985–4009 (2014).
4. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol*. **8**, 923–925 (2001).
5. wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. **47**, D520–D528 (2018).
6. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res*. **48**, D570–D578 (2020).
7. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
8. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys*. **37**, 289–316 (2008).
9. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
10. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
11. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Comput. Biol*. **13**, e1005324 (2017).
12. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).
13. Abriata, L. A., Tamò, G. E. & Dal Peraro, M. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins* **87**, 1100–1112 (2019).
14. Pearce, R. & Zhang, Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol*. **68**, 194–207 (2021).
15. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Topf, M. Critical assessment of techniques for protein structure prediction, fourteenth round. *CASP 14 Abstract Book* https://www.predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf (2020).
16. Brini, E., Simmerling, C. & Dill, K. Protein storytelling through physics. *Science* **370**, eaaz3041 (2020).
17. Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol*. **213**, 859–883 (1990).
18. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*. **234**, 779–815 (1993).
19. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protocols* **5**, 725–738 (2010).
20. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol*. **193**, 693–707 (1987).
21. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*. **7**, 349–358 (1994).
22. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
23. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
24. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
25. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–iv (1995).
26. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-round XIII. *Proteins* **87**, 1011–1020 (2019).
27. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
28. Tu, Z. & Bai, X. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*. **32**, 1744–1757 (2010).
29. Carreira, J., Agrawal, P., Fragkiadaki, K. & Malik, J. Human pose estimation with iterative error feedback. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4733–4742 (2016).
30. Mirabello, C. & Wallner, B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PLoS ONE* **14**, e0220182 (2019).
31. Huang, Z. et al. CCNet: criss-cross attention for semantic segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision* 603–612 (2019).
32. Hornak, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).
33. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. **31**, 3370–3374 (2003).
34. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
35. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10687–10698 (2020).
36. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res*. **45**, D170–D176 (2017).
37. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1, 4171–4186 (2019).
38. Rao, R. et al. MSA transformer. In *Proc. 38th International Conference on Machine Learning* PMLR 139, 8844–8856 (2021).
39. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* https://doi.org/10.1038/s41586-021-03828-1 (2021).
40. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol*. **20**, 681–697 (2019).
41. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol*. **30**, 1072–1080 (2012).
42. Qian, N. & Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol*. **202**, 865–884 (1988).
43. Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng*. **14**, 835–843 (2001).
44. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
45. Li, Y. et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Comput. Biol*. **17**, e1008865 (2021).
46. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
47. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst*. **8**, 292–301 (2019).
48. Senior, A. W. et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).
49. Ingraham, J., Riesselman, A. J., Sander, C. & Marks, D. S. Learning protein structure with a differentiable simulator. in *Proc. International Conference on Learning Representations* (2019).
50. Li, J. Universal transforming geometric network. Preprint at https://arxiv.org/abs/1908.00723 (2019).
51. Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach. Intell*. **3**, 601–609 (2021).
52. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (2017).
53. Wang, H. et al. Axial-deeplab: stand-alone axial-attention for panoptic segmentation. in *European Conference on Computer Vision* 108–126 (Springer, 2020).
54. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
55. Heinzinger, M. et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).
56. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
57. Pereira, J. et al. High-accuracy protein structure prediction in CASP14. *Proteins* https://doi.org/10.1002/prot.26171 (2021).
58. Gupta, M. et al. CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. Preprint at https://doi.org/10.1101/2021.05.10.443524 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Methods

Full algorithm details

Extensive explanations of the components and their motivations are available in Supplementary Methods 1.1–1.10, in addition, pseudocode is available in Supplementary Information Algorithms 1–32, network diagrams in Supplementary Figs. 1–8, input features in Supplementary Table 1 and additional details are provided in Supplementary Tables 2, 3. Training and inference details are provided in Supplementary Methods 1.11–1.12 and Supplementary Tables 4, 5.

IPA

The IPA module combines the pair representation, the single representation and the geometric representation to update the single representation (Supplementary Fig. 8). Each of these representations contributes affinities to the shared attention weights and then uses these weights to map its values to the output. The IPA operates in 3D space. Each residue produces query points, key points and value points in its local frame. These points are projected into the global frame using the backbone frame of the residue in which they interact with each other. The resulting points are then projected back into the local frame. The affinity computation in the 3D space uses squared distances and the coordinate transformations ensure the invariance of this module with respect to the global frame (see Supplementary Methods 1.8.2 'Invariant point attention (IPA)' for the algorithm, proof of invariance and a description of the full multi-head version). A related construction that uses classic geometric invariants to construct pairwise features in place of the learned 3D points has been applied to protein design[59].

In addition to the IPA, standard dot product attention is computed on the abstract single representation and a special attention on the pair representation. The pair representation augments both the logits and the values of the attention process, which is the primary way in which the pair representation controls the structure generation.

Inputs and data sources

Inputs to the network are the primary sequence, sequences from evolutionarily related proteins in the form of a MSA created by standard tools including jackhmmer[60] and HHBlits[61], and 3D atom coordinates of a small number of homologous structures (templates) where available. For both the MSA and templates, the search processes are tuned for high recall; spurious matches will probably appear in the raw MSA but this matches the training condition of the network.

One of the sequence databases used, Big Fantastic Database (BFD), was custom-made and released publicly (see 'Data availability') and was used by several CASP teams. BFD is one of the largest publicly available collections of protein families. It consists of 65,983,866 families represented as MSAs and hidden Markov models (HMMs) covering 2,204,359,010 protein sequences from reference databases, metagenomes and metatranscriptomes.

BFD was built in three steps. First, 2,423,213,294 protein sequences were collected from UniProt (Swiss-Prot&TrEMBL, 2017-11)[62], a soil reference protein catalogue and the marine eukaryotic reference catalogue[7], and clustered to 30% sequence identity, while enforcing a 90% alignment coverage of the shorter sequences using MMseqs2/Linclust[63]. This resulted in 345,159,030 clusters. For computational efficiency, we removed all clusters with less than three members, resulting in 61,083,719 clusters. Second, we added 166,510,624 representative protein sequences from Metaclust NR (2017-05; discarding all sequences shorter than 150 residues)[63] by aligning them against the cluster representatives using MMseqs2[64]. Sequences that fulfilled the sequence identity and coverage criteria were assigned to the best scoring cluster. The remaining 25,347,429 sequences that could not be assigned were clustered separately and added as new clusters, resulting in the final clustering. Third, for each of the clusters, we computed an MSA using FAMSA[65] and computed the HMMs following the Uniclust HH-suite database protocol[36].

The following versions of public datasets were used in this study. Our models were trained on a copy of the PDB[5] downloaded on 28 August 2019. For finding template structures at prediction time, we used a copy of the PDB downloaded on 14 May 2020, and the PDB70[66] clustering database downloaded on 13 May 2020. For MSA lookup at both training and prediction time, we used Uniref90[67] v.2020_01, BFD, Uniclust30[36] v.2018_08 and MGnify[6] v.2018_12. For sequence distillation, we used Uniclust30[36] v.2018_08 to construct a distillation structure dataset. Full details are provided in Supplementary Methods 1.2.

For MSA search on BFD + Uniclust30, and template search against PDB70, we used HHBlits[61] and HHSearch[66] from hh-suite v.3.0-beta.3 (version 14/07/2017). For MSA search on Uniref90 and clustered MGnify, we used jackhmmer from HMMER3[68]. For constrained relaxation of structures, we used OpenMM v.7.3.1[69] with the Amber99sb force field[32]. For neural network construction, running and other analyses, we used TensorFlow[70], Sonnet[71], NumPy[72], Python[73] and Colab[74].

To quantify the effect of the different sequence data sources, we re-ran the CASP14 proteins using the same models but varying how the MSA was constructed. Removing BFD reduced the mean accuracy by 0.4 GDT, removing Mgnify reduced the mean accuracy by 0.7 GDT, and removing both reduced the mean accuracy by 6.1 GDT. In each case, we found that most targets had very small changes in accuracy but a few outliers had very large (20+ GDT) differences. This is consistent with the results in Fig. 5a in which the depth of the MSA is relatively unimportant until it approaches a threshold value of around 30 sequences when the MSA size effects become quite large. We observe mostly overlapping effects between inclusion of BFD and Mgnify, but having at least one of these metagenomics databases is very important for target classes that are poorly represented in UniRef, and having both was necessary to achieve full CASP accuracy.

Training regimen

To train, we use structures from the PDB with a maximum release date of 30 April 2018. Chains are sampled in inverse proportion to cluster size of a 40% sequence identity clustering. We then randomly crop them to 256 residues and assemble into batches of size 128. We train the model on Tensor Processing Unit (TPU) v3 with a batch size of 1 per TPU core, hence the model uses 128 TPU v3 cores. The model is trained until convergence (around 10 million samples) and further fine-tuned using longer crops of 384 residues, larger MSA stack and reduced learning rate (see Supplementary Methods 1.11 for the exact configuration). The initial training stage takes approximately 1 week, and the fine-tuning stage takes approximately 4 additional days.

The network is supervised by the FAPE loss and a number of auxiliary losses. First, the final pair representation is linearly projected to a binned distance distribution (distogram) prediction, scored with a cross-entropy loss. Second, we use random masking on the input MSAs and require the network to reconstruct the masked regions from the output MSA representation using a BERT-like loss[37]. Third, the output single representations of the structure module are used to predict binned per-residue lDDT-Cα values. Finally, we use an auxiliary side-chain loss during training, and an auxiliary structure violation loss during fine-tuning. Detailed descriptions and weighting are provided in the Supplementary Information.

An initial model trained with the above objectives was used to make structure predictions for a Uniclust dataset of 355,993 sequences with the full MSAs. These predictions were then used to train a final model with identical hyperparameters, except for sampling examples 75% of the time from the Uniclust prediction set, with sub-sampled MSAs, and 25% of the time from the clustered PDB set.

We train five different models using different random seeds, some with templates and some without, to encourage diversity in the predictions (see Supplementary Table 5 and Supplementary Methods 1.12.1

for details). We also fine-tuned these models after CASP14 to add a pTM prediction objective (Supplementary Methods 1.9.7) and use the obtained models for Fig. 2d.

Inference regimen

We inference the five trained models and use the predicted confidence score to select the best model per target.

Using our CASP14 configuration for AlphaFold, the trunk of the network is run multiple times with different random choices for the MSA cluster centres (see Supplementary Methods 1.11.2 for details of the ensembling procedure). The full time to make a structure prediction varies considerably depending on the length of the protein. Representative timings for the neural network using a single model on V100 GPU are 4.8 min with 256 residues, 9.2 min with 384 residues and 18 h at 2,500 residues. These timings are measured using our open-source code, and the open-source code is notably faster than the version we ran in CASP14 as we now use the XLA compiler[75].

Since CASP14, we have found that the accuracy of the network without ensembling is very close or equal to the accuracy with ensembling and we turn off ensembling for most inference. Without ensembling, the network is 8× faster and the representative timings for a single model are 0.6 min with 256 residues, 1.1 min with 384 residues and 2.1 h with 2,500 residues.

Inferencing large proteins can easily exceed the memory of a single GPU. For a V100 with 16 GB of memory, we can predict the structure of proteins up to around 1,300 residues without ensembling and the 256- and 384-residue inference times are using the memory of a single GPU. The memory usage is approximately quadratic in the number of residues, so a 2,500-residue protein involves using unified memory so that we can greatly exceed the memory of a single V100. In our cloud setup, a single V100 is used for computation on a 2,500-residue protein but we requested four GPUs to have sufficient memory.

Searching genetic sequence databases to prepare inputs and final relaxation of the structures take additional central processing unit (CPU) time but do not require a GPU or TPU.

Metrics

The predicted structure is compared to the true structure from the PDB in terms of lDDT metric[34], as this metric reports the domain accuracy without requiring a domain segmentation of chain structures. The distances are either computed between all heavy atoms (lDDT) or only the Cα atoms to measure the backbone accuracy (lDDT-Cα). As lDDT-Cα only focuses on the Cα atoms, it does not include the penalty for structural violations and clashes. Domain accuracies in CASP are reported as GDT[33] and the TM-score[27] is used as a full chain global superposition metric.

We also report accuracies using the r.m.s.d.$_{95}$ (Cα r.m.s.d. at 95% coverage). We perform five iterations of (1) a least-squares alignment of the predicted structure and the PDB structure on the currently chosen Cα atoms (using all Cα atoms in the first iteration); (2) selecting the 95% of Cα atoms with the lowest alignment error. The r.m.s.d. of the atoms chosen for the final iterations is the r.m.s.d.$_{95}$. This metric is more robust to apparent errors that can originate from crystal structure artefacts, although in some cases the removed 5% of residues will contain genuine modelling errors.

Test set of recent PDB sequences

For evaluation on recent PDB sequences (Figs. 2a–d, 4a, 5a), we used a copy of the PDB downloaded 15 February 2021. Structures were filtered to those with a release date after 30 April 2018 (the date limit for inclusion in the training set for AlphaFold). Chains were further filtered to remove sequences that consisted of a single amino acid as well as sequences with an ambiguous chemical component at any residue position. Exact duplicates were removed, with the chain with the most resolved Cα atoms used as the representative sequence. Subsequently, structures with less than 16 resolved residues, with unknown residues or solved by NMR methods were removed. As the PDB contains many near-duplicate sequences, the chain with the highest resolution was selected from each cluster in the PDB 40% sequence clustering of the data. Furthermore, we removed all sequences for which fewer than 80 amino acids had the alpha carbon resolved and removed chains with more than 1,400 residues. The final dataset contained 10,795 protein sequences.

The procedure for filtering the recent PDB dataset based on prior template identity was as follows. Hmmsearch was run with default parameters against a copy of the PDB SEQRES fasta downloaded 15 February 2021. Template hits were accepted if the associated structure had a release date earlier than 30 April 2018. Each residue position in a query sequence was assigned the maximum identity of any template hit covering that position. Filtering then proceeded as described in the individual figure legends, based on a combination of maximum identity and sequence coverage.

The MSA depth analysis was based on computing the normalized number of effective sequences (N_{eff}) for each position of a query sequence. Per-residue N_{eff} values were obtained by counting the number of non-gap residues in the MSA for this position and weighting the sequences using the N_{eff} scheme[76] with a threshold of 80% sequence identity measured on the region that is non-gap in either sequence.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All input data are freely available from public sources.

Structures from the PDB were used for training and as templates (https://www.wwpdb.org/ftp/pdb-ftp-sites; for the associated sequence data and 40% sequence clustering see also https://ftp.wwpdb.org/pub/pdb/derived_data/ and https://cdn.rcsb.org/resources/sequence/clusters/bc-40.out). Training used a version of the PDB downloaded 28 August 2019, while the CASP14 template search used a version downloaded 14 May 2020. The template search also used the PDB70 database, downloaded 13 May 2020 (https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/).

We show experimental structures from the PDB with accession numbers 6Y4F[77], 6YJ1[78], 6VR4[79], 6SK0[80], 6FES[81], 6W6W[82], 6T1Z[83] and 7JTL[84].

For MSA lookup at both the training and prediction time, we used UniRef90 v.2020_01 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/), BFD (https://bfd.mmseqs.com), Uniclust30 v.2018_08 (https://wwwuser.gwdg.de/~compbiol/uniclust/2018_08/) and MGnify clusters v.2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/). Uniclust30 v.2018_08 was also used as input for constructing a distillation structure dataset.

Code availability

Source code for the AlphaFold model, trained weights and inference script are available under an open-source license at https://github.com/deepmind/alphafold.

Neural networks were developed with TensorFlow v.1 (https://github.com/tensorflow/tensorflow), Sonnet v.1 (https://github.com/deepmind/sonnet), JAX v.0.1.69 (https://github.com/google/jax/) and Haiku v.0.0.4 (https://github.com/deepmind/dm-haiku). The XLA compiler is bundled with JAX and does not have a separate version number.

For MSA search on BFD+Uniclust30, and for template search against PDB70, we used HHBlits and HHSearch from hh-suite v.3.0-beta.3 release 14/07/2017 (https://github.com/soedinglab/hh-suite). For MSA search on UniRef90 and clustered MGnify, we used jackhmmer from

HMMER v.3.3 (http://eddylab.org/software/hmmer/). For constrained relaxation of structures, we used OpenMM v.7.3.1 (https://github.com/openmm/openmm) with the Amber99sb force field.

Construction of BFD used MMseqs2 v.925AF (https://github.com/soedinglab/MMseqs2) and FAMSA v.1.2.5 (https://github.com/refresh-bio/FAMSA).

Data analysis used Python v.3.6 (https://www.python.org/), NumPy v.1.16.4 (https://github.com/numpy/numpy), SciPy v.1.2.1 (https://www.scipy.org/), seaborn v.0.11.1 (https://github.com/mwaskom/seaborn), Matplotlib v.3.3.4 (https://github.com/matplotlib/matplotlib), bokeh v.1.4.0 (https://github.com/bokeh/bokeh), pandas v.1.1.5 (https://github.com/pandas-dev/pandas), plotnine v.0.8.0 (https://github.com/has2k1/plotnine), statsmodels v.0.12.2 (https://github.com/statsmodels/statsmodels) and Colab (https://research.google.com/colaboratory). TM-align v.20190822 (https://zhanglab.dcmb.med.umich.edu/TM-align/) was used for computing TM-scores. Structure visualizations were created in Pymol v.2.3.0 (https://github.com/schrodinger/pymol-open-source).

59. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. in *Proc. 33rd Conference on Neural Information Processing Systems* (2019).
60. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
61. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
62. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).
63. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
64. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
65. Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964 (2016).
66. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
67. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
68. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
69. Eastman, P. et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).
70. Ashish, A. M. A. et al. TensorFlow: large-scale machine learning on heterogeneous systems. Preprint at https://arxiv.org/abs/1603.04467 (2015).
71. Reynolds, M. et al. Open sourcing Sonnet – a new library for constructing neural networks. *DeepMind* https://deepmind.com/blog/open-sourcing-sonnet/ (7 April 2017).
72. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
73. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, 2009).
74. Bisong, E. in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* 59–64 (Apress, 2019).
75. TensorFlow. XLA: Optimizing Compiler for TensorFlow. https://www.tensorflow.org/xla (2018).
76. Wu, T., Hou, J., Adhikari, B. & Cheng, J. Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics* **36**, 1091–1098 (2020).
77. Jiang, W. et al. MrpH, a new class of metal-binding adhesin, requires zinc to mediate biofilm formation. *PLoS Pathog.* **16**, e1008707 (2020).
78. Dunne, M., Ernst, P., Sobieraj, A., Pluckthun, A. & Loessner, M. J. The M23 peptidase domain of the Staphylococcal phage 2638A endolysin. *PDB* https://doi.org/10.2210/pdb6YJ1/pdb (2020).
79. Drobysheva, A. V. et al. Structure and function of virion RNA polymerase of a crAss-like phage. *Nature* **589**, 306–309 (2021).
80. Flaugnatti, N. et al. Structural basis for loading and inhibition of a bacterial T6SS phospholipase effector by the VgrG spike. *EMBO J.* **39**, e104129 (2020).
81. ElGamacy, M. et al. An interface-driven design strategy yields a novel, corrugated protein architecture. *ACS Synth. Biol.* **7**, 2226–2235 (2018).
82. Lim, C. J. et al. The structure of human CST reveals a decameric assembly bound to telomeric DNA. *Science* **368**, 1081–1085 (2020).
83. Debruycker, V. et al. An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nat. Struct. Mol. Biol.* **27**, 829–835 (2020).
84. Flower, T. G. et al. Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc. Natl Acad. Sci. USA* **118**, e2021785118 (2021).

Acknowledgements We thank A. Rrustemi, A. Gu, A. Guseynov, B. Hechtman, C. Beattie, C. Jones, C. Donner, E. Parisotto, E. Elsen, F. Popovici, G. Necula, H. Maclean, J. Menick, J. Kirkpatrick, J. Molloy, J. Yim, J. Stanway, K. Simonyan, L. Sifre, L. Martens, M. Johnson, M. O'Neill, N. Antropova, R. Hadsell, S. Blackwell, S. Das, S. Hou, S. Gouws, S. Wheelwright, T. Hennigan, T. Ward, Z. Wu, Ž. Avsec and the Research Platform Team for their contributions; M. Mirdita for his help with the datasets; M. Piovesan-Forster, A. Nelson and R. Kemp for their help managing the project; the JAX, TensorFlow and XLA teams for detailed support and enabling machine learning models of the complexity of AlphaFold; our colleagues at DeepMind, Google and Alphabet for their encouragement and support; and J. Moult and the CASP14 organizers, and the experimentalists whose structures enabled the assessment. M.S. acknowledges support from the National Research Foundation of Korea grant (2019R1A6A1A10073437, 2020M3A9G7103933) and the Creative-Pioneering Researchers Program through Seoul National University.

Author contributions J.J. and D.H. led the research. J.J., R.E., A. Pritzel, M.F., O.R., R.B., A. Potapenko, S.A.A.K., B.R.-P., J.A., M.P., T. Berghammer and O.V. developed the neural network architecture and training. T.G., A.Ž., K.T., R.B., A.B., R.E., A.J.B., A.C., S.N., R.J., D.R., M.Z. and S.B. developed the data, analytics and inference systems. D.H., K.K., P.K., C.M. and E.C. managed the research. T.G. led the technical platform. P.K., A.W.S., K.K., O.V., D.S., S.P. and T. Back contributed technical advice and ideas. M.S. created the BFD genomics database and provided technical assistance on HHBlits. D.H., R.E., A.W.S. and K.K. conceived the AlphaFold project. J.J., R.E. and A.W.S. conceived the end-to-end approach. J.J., A. Pritzel, O.R., A. Potapenko, R.E., M.F., T.G., K.T., C.M. and D.H. wrote the paper.

Competing interests J.J., R.E., A. Pritzel, T.G., M.F., O.R., R.B., A.B., S.A.A.K., D.R. and A.W.S. have filed non-provisional patent applications 16/701,070 and PCT/EP2020/084238, and provisional patent applications 63/107,362, 63/118,917, 63/118,918, 63/118,921 and 63/118,919, each in the name of DeepMind Technologies Limited, each pending, relating to machine learning for predicting protein structures. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03819-2.

Correspondence and requests for materials should be addressed to J.J. or D.H.

Peer review information *Nature* thanks Mohammed AlQuraishi, Charlotte Deane and Yang Zhang for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

https://doi.org/10.1038/s41586-024-07487-w

Received: 19 December 2023

Accepted: 29 April 2024

Published online: 8 May 2024

Open access

Check for updates

Josh Abramson[1,7], Jonas Adler[1,7], Jack Dunger[1,7], Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], Olaf Ronneberger[1,7], Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick[2], Sebastian W. Bodenstein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'Neill[1], David Reiman[1], Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arvaniti[3], Charles Beattie[3], Ottavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles Congreve[4], Alexander I. Cowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabian B. Fuchs[3], Hannah Gladman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], Kuba Perlin[3], Anna Potapenko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Ashok Thillaisundaram[3], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], Michal Zielinski[3], Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8]✉, Demis Hassabis[1,2,8]✉ & John M. Jumper[1,8]✉

The introduction of AlphaFold 2[1] has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy over many previous specialized tools: far greater accuracy for protein–ligand interactions compared with state-of-the-art docking tools, much higher accuracy for protein–nucleic acid interactions compared with nucleic-acid-specific predictors and substantially higher antibody–antigen prediction accuracy compared with AlphaFold-Multimer v.2.3[7,8]. Together, these results show that high-accuracy modelling across biomolecular space is possible within a single unified deep-learning framework.

Accurate models of biological complexes are critical to our understanding of cellular functions and for the rational design of therapeutics[2–4,9]. Enormous progress has been achieved in protein structure prediction with the development of AlphaFold[1], and the field has grown tremendously with a number of later methods that build on the ideas and techniques of AlphaFold 2 (AF2)[10–12]. Almost immediately after AlphaFold became available, it was shown that simple input modifications would enable surprisingly accurate protein interaction predictions[13–15] and that training AF2 specifically for protein interaction prediction yielded a highly accurate system[7].

These successes lead to the question of whether it is possible to accurately predict the structure of complexes containing a much wider range of biomolecules, including ligands, ions, nucleic acids and modified residues, within a deep-learning framework. A wide range of predictors for various specific interaction types has been developed[16–28], as well as one generalist method developed concurrently with the present work[29], but the accuracy of such deep-learning attempts has been mixed and often below that of physics-inspired methods[30,31]. Almost all of these methods are also highly specialized to particular interaction types and cannot predict the structure of general biomolecular complexes containing many types of entities.

Here we present AlphaFold 3 (AF3)—a model that is capable of high-accuracy prediction of complexes containing nearly all molecular types present in the Protein Data Bank[32] (PDB) (Fig. 1a,b). In all but one category, it achieves a substantially higher performance than strong methods that specialize in just the given task (Fig. 1c and Extended Data Table 1), including higher accuracy at protein structure and the structure of protein–protein interactions.

This is achieved by a substantial evolution of the AF2 architecture and training procedure (Fig. 1d) both to accommodate more general chemical structures and to improve the data efficiency of learning. The system reduces the amount of multiple-sequence alignment (MSA) processing by replacing the AF2 evoformer with the simpler pairformer module (Fig. 2a). Furthermore it directly predicts the raw atom coordinates with a diffusion module, replacing the AF2 structure module that operated on amino-acid-specific frames and side-chain torsion angles (Fig. 2b). The multiscale nature of the diffusion process (low noise levels induce the network to improve local structure) also enable us to eliminate stereochemical losses and most special handling of bonding patterns in the network, easily accommodating arbitrary chemical components.

[1]Core Contributor, Google DeepMind, London, UK. [2]Core Contributor, Isomorphic Labs, London, UK. [3]Google DeepMind, London, UK. [4]Isomorphic Labs, London, UK. [5]Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA, USA. [6]Department of Computer Science, Princeton University, Princeton, NJ, USA. [7]These authors contributed equally: Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore. [8]These authors jointly supervised this work: Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, John M. Jumper. ✉e-mail: jaderberg@isomorphiclabs.com; dhcontact@google.com; jumper@google.com

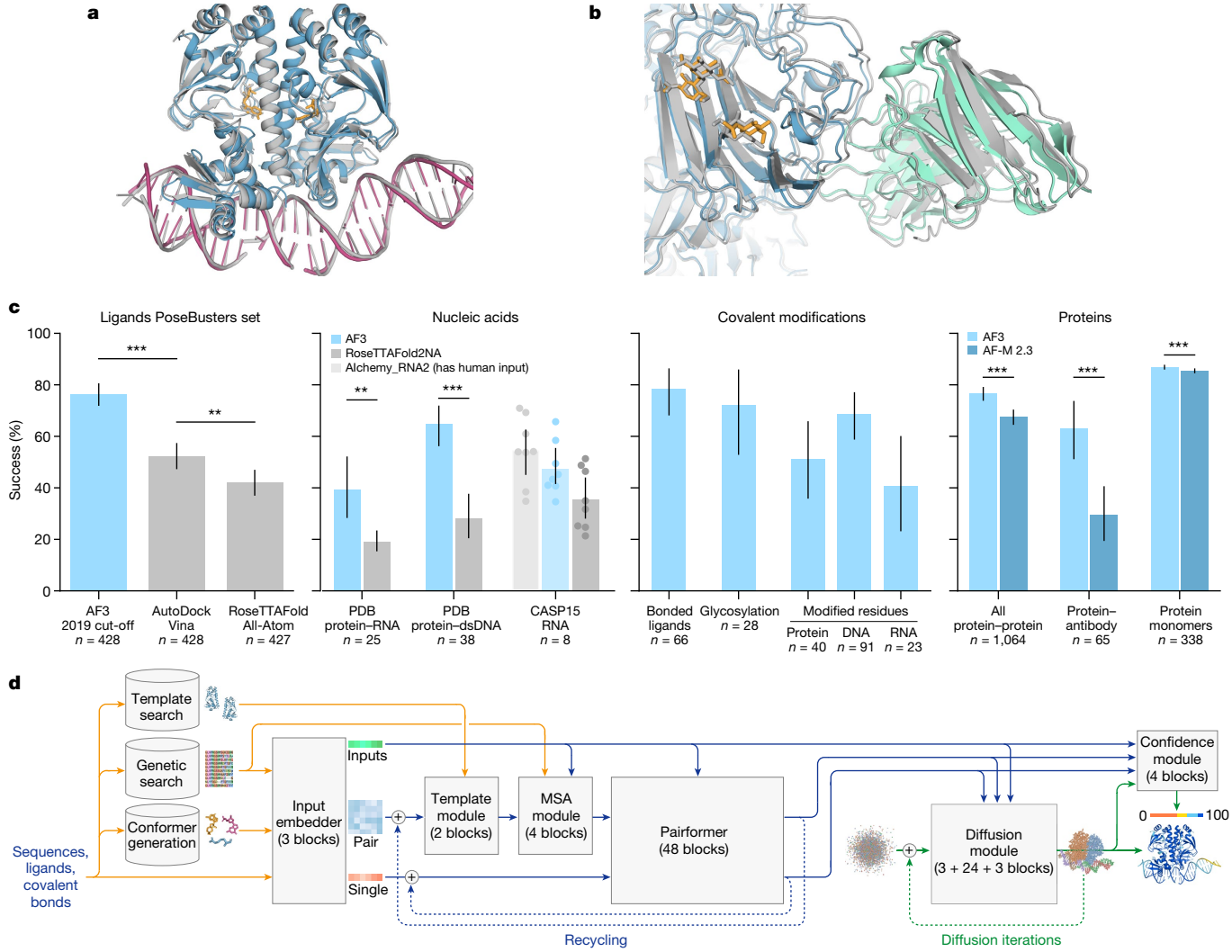


Fig. 1 | AF3 accurately predicts structures across biomolecular complexes. **a**,**b**, Example structures predicted using AF3. **a**, Bacterial CRP/FNR family transcriptional regulator protein bound to DNA and cGMP (PDB 7PZB; full-complex LDDT[47], 82.8; global distance test (GDT)[48], 90.1). **b**, Human coronavirus OC43 spike protein, 4,665 residues, heavily glycosylated and bound by neutralizing antibodies (PDB 7PNM; full-complex LDDT, 83.0; GDT, 83.1). **c**, AF3 performance on PoseBusters (v.1, August 2023 release), our recent PDB evaluation set and CASP15 RNA. Metrics are as follows: percentage of pocket-aligned ligand r.m.s.d. < 2 Å for ligands and covalent modifications; interface LDDT for protein–nucleic acid complexes; LDDT for nucleic acid and protein monomers; and percentage DockQ > 0.23 for protein–protein and protein–antibody interfaces. All scores are reported from the top confidence-ranked sample out of five model seeds (each with five diffusion samples), except for protein–antibody scores, which were ranked across 1,000 model seeds for both models (each AF3 seed with five diffusion samples). Sampling and ranking details are provided in the Methods. For ligands, *n* indicates the number of targets; for nucleic acids, *n* indicates the number of structures; for modifications, *n* indicates the number of clusters; and for proteins, *n* indicates the number of clusters. The bar height indicates the mean; error bars indicate exact binomial distribution 95% confidence intervals for PoseBusters and by 10,000 bootstrap resamples for all others. Significance levels were calculated using two-sided Fisher's exact tests for PoseBusters and using two-sided Wilcoxon signed-rank tests for all others; $^{***}P < 0.001$, $^{**}P < 0.01$. Exact *P* values (from left to right) are as follows: 2.27×10^{-13}, 2.57×10^{-3}, 2.78×10^{-3}, 7.28×10^{-12}, 1.81×10^{-18}, 6.54×10^{-5} and 1.74×10^{-34}. AF-M 2.3, AlphaFold-Multimer v.2.3; dsDNA, double-stranded DNA. **d**, AF3 architecture for inference. The rectangles represent processing modules and the arrows show the data flow. Yellow, input data; blue, abstract network activations; green, output data. The coloured balls represent physical atom coordinates.

Network architecture and training

The overall structure of AF3 (Fig. 1d and Supplementary Methods 3) echoes that of AF2, with a large trunk evolving a pairwise representation of the chemical complex followed by a structure module that uses the pairwise representation to generate explicit atomic positions, but there are large differences in each major component. These modifications were driven both by the need to accommodate a wide range of chemical entities without excessive special casing and by observations of AF2 performance with different modifications. Within the trunk, MSA processing is substantially de-emphasized, with a much smaller and simpler MSA embedding block (Supplementary Methods 3.3). Compared with the original evoformer from AF2, the number of blocks is reduced to four, the processing of the MSA representation uses an inexpensive pair-weighted averaging and only the pair representation is used for later processing steps. The 'pairformer' (Fig. 2a and Supplementary Methods 3.6) replaces the evoformer of AF2 as the dominant processing block. It operates only on the pair representation and the single representation; the MSA representation is not retained and all information passes through the pair representation. The pair processing and the number of blocks (48) is largely unchanged from AF2. The resulting pair and single representation together with the input representation are passed to the new diffusion module (Fig. 2b) that replaces the structure module of AF2.

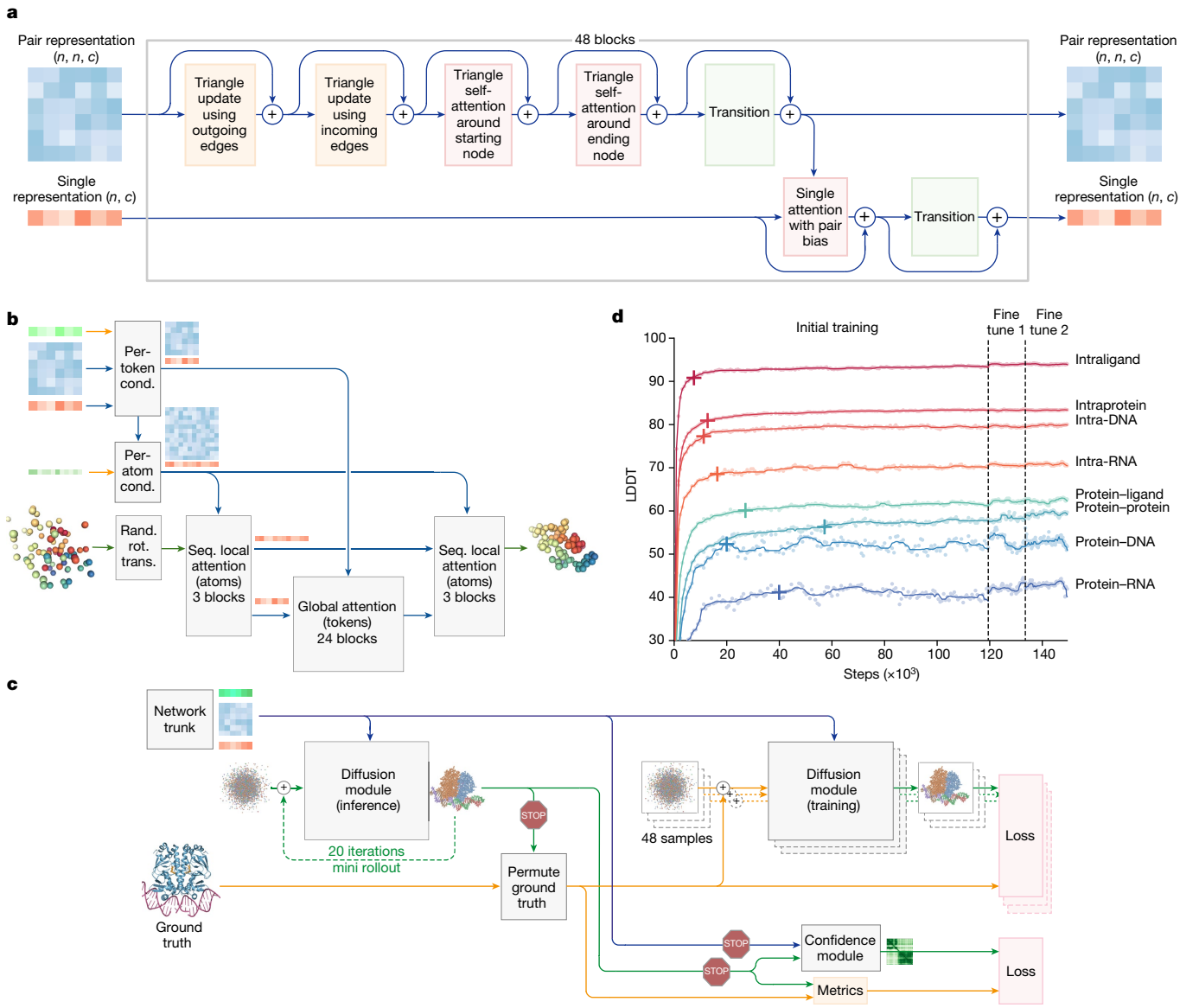


Fig. 2 | Architectural and training details. a, The pairformer module. Input and output: pair representation with dimension (n, n, c) and single representation with dimension (n, c). n is the number of tokens (polymer residues and atoms); c is the number of channels (128 for the pair representation, 384 for the single representation). Each of the 48 blocks has an independent set of trainable parameters. **b**, The diffusion module. Input: coarse arrays depict per-token representations (green, inputs; blue, pair; red, single). Fine arrays depict per-atom representations. The coloured balls represent physical atom coordinates. Cond., conditioning; rand. rot. trans., random rotation and translation; seq., sequence. **c**, The training set-up (distogram head omitted) starting from the end of the network trunk. The coloured arrays show activations from the network trunk (green, inputs; blue, pair; red, single). The blue arrows show abstract activation arrays. The yellow arrows show ground-truth data. The green arrows show predicted data. The stop sign represents stopping of the gradient. Both depicted diffusion modules share weights. **d**, Training curves for initial training and fine-tuning stages, showing the LDDT on our evaluation set as a function of optimizer steps. The scatter plot shows the raw datapoints and the lines show the smoothed performance using a median filter with a kernel width of nine datapoints. The crosses mark the point at which the smoothed performance reaches 97% of its initial training maximum.

The diffusion module (Fig. 2b and Supplementary Methods 3.7) operates directly on raw atom coordinates, and on a coarse abstract token representation, without rotational frames or any equivariant processing. We had observed in AF2 that removing most of the complexity of the structure module had only a modest effect on the prediction accuracy, and maintaining the backbone frame and side-chain torsion representation add quite a bit of complexity for general molecular graphs. Similarly AF2 required carefully tuned stereochemical violation penalties during training to enforce chemical plausibility of the resulting structures. We use a relatively standard diffusion approach[33] in which the diffusion model is trained to receive 'noised' atomic coordinates and then predict the true coordinates. This task requires the network to learn protein structure at a variety of length scales, whereby the denoising task at small noise emphasizes understanding very local stereochemistry and the denoising task at high noise emphasizes the large-scale structure of the system. At the inference time, random noise is sampled and then recurrently denoised to produce a final structure. Importantly, this is a generative training procedure that produces a distribution of answers. This means that, for each answer, the local structure will be sharply defined (for example, side-chain bond geometry) even when the network is uncertain about the positions. For this reason, we are able to avoid both torsion-based parametrizations of the residues and violation losses on the structure, while handling the full complexity of general ligands. Similarly to some recent work[34],

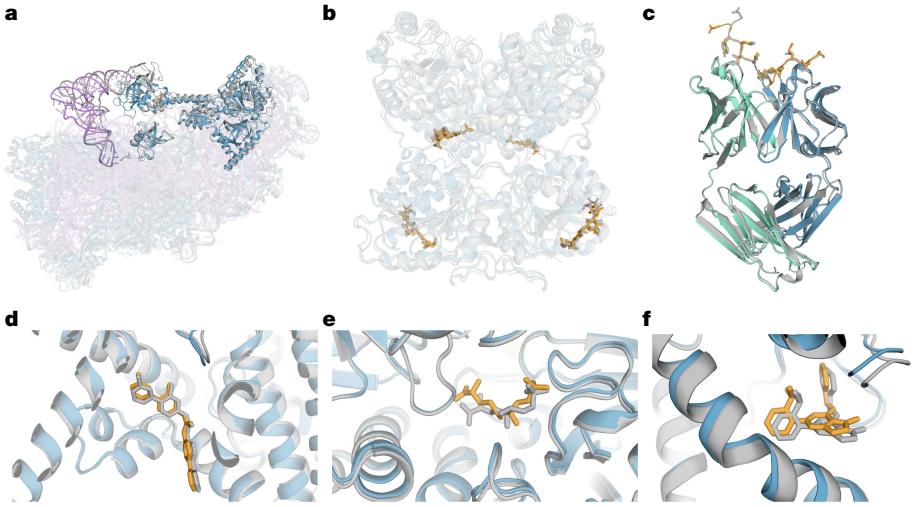


Fig. 3 | Examples of predicted complexes. Selected structure predictions from AF3. Predicted protein chains are shown in blue (predicted antibody in green), predicted ligands and glycans in orange, predicted RNA in purple and the ground truth is shown in grey. **a**, Human 40S small ribosomal subunit (7,663 residues) including 18S ribosomal RNA and Met-tRNA$_i^{Met}$ (opaque purple) in a complex with translation initiation factors eIF1A and eIF5B (opaque blue; PDB 7TQL; full-complex LDDT, 87.7; GDT, 86.9). **b**, The glycosylated globular portion of an EXTL3 homodimer (PDB 7AU2; mean pocket-aligned r.m.s.d., 1.10 Å). **c**, Mesothelin C-terminal peptide bound to the monoclonal antibody 15B6 (PDB 7U8C; DockQ, 0.85). **d**, LGK974, a clinical-stage inhibitor, bound to PORCN in a complex with the WNT3A peptide (PDB 7URD; ligand r.m.s.d., 1.00 Å). **e**, (5S,6S)-O7-sulfo DADH bound to the AziU3/U2 complex with a novel fold (PDB 7WUX; ligand r.m.s.d., 1.92 Å). **f**, Analogue of NIH-12848 bound to an allosteric site of PI5P4Kγ (PDB 7QIE; ligand r.m.s.d., 0.37 Å).

we find that no invariance or equivariance with respect to global rotations and translation of the molecule are required in the architecture and we therefore omit them to simplify the machine learning architecture.

The use of a generative diffusion approach comes with some technical challenges that we needed to address. The biggest issue is that generative models are prone to hallucination[35], whereby the model may invent plausible-looking structure even in unstructured regions. To counteract this effect, we use a cross-distillation method in which we enrich the training data with structures predicted by AlphaFold-Multimer (v.2.3)[7,8]. In these structures, unstructured regions are typically represented by long extended loops instead of compact structures, and training on them 'teaches' AF3 to mimic this behaviour. This cross-distillation greatly reduced the hallucination behaviour of AF3 (Extended Data Fig. 1 for disorder prediction results on the CAID 2[36] benchmark set).

We also developed confidence measures that predict the atom-level and pairwise errors in our final structures. In AF2, this was done directly by regressing the error in the output of the structure module during training. However, this procedure is not applicable to diffusion training, as only a single step of the diffusion is trained instead of a full-structure generation (Fig. 2c). To remedy this, we developed a diffusion 'rollout' procedure for the full-structure prediction generation during training (using a larger step size than normal; Fig. 2c (mini-rollout)). This predicted structure is then used to permute the symmetric ground-truth chains and ligands, and to compute the performance metrics to train the confidence head. The confidence head uses the pairwise representation to predict a modified local distance difference test (pLDDT) and a predicted aligned error (PAE) matrix as in AF2, as well as a distance error matrix (PDE), which is the error in the distance matrix of the predicted structure as compared to the true structure (details are provided in Supplementary Methods 4.3).

Figure 2d shows that, during initial training, the model learns quickly to predict the local structures (all intrachain metrics go up quickly and reach 97% of the maximum performance within the first 20,000 training steps), while the model needs considerably longer to learn the global constellation (the interface metrics go up slowly and protein–protein interface LDDT passes the 97% bar only after 60,000 steps). During AF3 development, we observed that some model abilities topped out relatively early and started to decline (most likely due to overfitting to the limited number of training samples for this capability), while other abilities were still undertrained. We addressed this by increasing or decreasing the sampling probability for the corresponding training sets (Supplementary Methods 2.5.1) and by performing early stopping using a weighted average of all of the above metrics and some additional metrics to select the best model checkpoint (Supplementary Table 7). The fine-tuning stages with the larger crop sizes improve the model on all metrics with an especially high uplift on protein–protein interfaces (Extended Data Fig. 2).

Accuracy across complex types

AF3 can predict structures from input polymer sequences, residue modifications and ligand SMILES (simplified molecular-input line-entry system). In Fig. 3 we show a selection of examples highlighting the ability of the model to generalize to a number of biologically important and therapeutically relevant modalities. In selecting these examples, we considered novelty in terms of the similarity of individual chains and interfaces to the training set (additional information is provided in Supplementary Methods 8.1).

We evaluated the performance of the system on recent interface-specific benchmarks for each complex type (Fig. 1c and Extended Data Table 1). Performance on protein–ligand interfaces was evaluated on the PoseBusters benchmark set, which is composed of 428 protein–ligand structures released to the PDB in 2021 or later. As our standard training cut-off date is in 2021, we trained a separate AF3 model with an earlier training-set cutoff (Methods). Accuracy on the PoseBusters set is reported as the percentage of protein–ligand pairs with pocket-aligned ligand root mean squared deviation (r.m.s.d.) of less than 2 Å. The baseline models come in two categories: those that use only protein sequence and ligand SMILES as an input and those that additionally leak information from the solved protein–ligand test structure. Traditional docking methods use the latter privileged information, even though that information would not be available in real-world use cases. Even so, AF3 greatly outperforms classical docking tools such as Vina[37,38] even while not using any structural inputs (Fisher's exact test, $P = 2.27 \times 10^{-13}$) and greatly outperforms all other true blind docking

like RoseTTAFold All-Atom ($P = 4.45 \times 10^{-25}$). Extended Data Fig. 3 shows three examples in which AF3 achieves accurate predictions but docking tools Vina and Gold do not[37]. PoseBusters analysis was performed using a training cut-off of 30 September 2019 for AF3 to ensure that the model was not trained on any PoseBusters structures. To compare with the RoseTTAFold All-Atom results, we used PoseBusters version 1. Version 2 (crystal contacts removed from the benchmark set) results including quality metrics are shown in Extended Data Fig. 4b–f and Extended Data Table 1. We use multiple seeds to ensure correct chirality and avoid slight protein–ligand clashing (as opposed to a method like diffusion guidance to enforce) but we are typically able to produce high-quality stereochemistry. Separately, we also train a version of AF3 that receives the 'pocket information' as used in some recent deep-learning work[24,26] (the results are shown in Extended Data Fig. 4a).

AF3 predicts protein–nucleic complexes and RNA structures with higher accuracy than RoseTTAFold2NA[15] (Fig. 1c (second plot)). As RoseTTAFold2NA is validated only on structures below 1,000 residues, we use only structures below 1,000 residues from our recent PDB evaluation set for this comparison (Methods). AF3 is able to predict protein–nucleic structures with thousands of residues, an example of which is shown in Fig. 3a. Note that we do not compare directly to RoseTTAFold All-Atom, but benchmarks indicate that RoseTTAFold All-Atom is slightly less accurate than RoseTTAFold2NA for nucleic acid predictions[29].

We also evaluated AF3 performance on the ten publicly available Critical Assessment of Structure Prediction 15 (CASP15) RNA targets: we achieve a higher average performance than RoseTTAFold2NA and AIchemy_RNA[27] (the best AI-based submission in CASP15[18,31]) on the respective common subsets of our and their predictions (detailed results are shown in Extended Data Fig. 5a). We did not reach the performance of the best human-expert-aided CASP15 submission AIchemy_RNA2[39] (Fig. 1c (centre left)). Owing to limited dataset sizes, we do not report significance test statistics here. Further analysis of the accuracy of predicting nucleic acids alone (without proteins) is shown in Extended Data Fig. 5b.

Covalent modifications (bonded ligands, glycosylation, and modified protein residues and nucleic acid bases) are also accurately predicted by AF3 (Fig. 1c (centre right)). Modifications include those to any polymer residue (protein, RNA or DNA). We report accuracy as the percentage of successful predictions (pocket r.m.s.d. < 2 Å). We apply quality filters to the bonded ligands and glycosylation dataset (as does PoseBusters): we include only ligands with high-quality experimental data (ranking_model_fit > 0.5, according to the RCSB structure validation report, that is, X-ray structures with a model quality above the median). As with the PoseBusters set, the bonded ligands and glycosylation datasets are not filtered by homology to the training dataset. Filtering on the basis of the bound polymer chain homology (using polymer template similarity < 40) yielded only five clusters for bonded ligands and seven clusters for glycosylation. We exclude multi-residue glycans here because the RCSB validation report does not provide a ranking_model_fit value for them. The percentage of successful predictions (pocket r.m.s.d. < 2 Å) for multi-residue glycans on all-quality experimental data is 42.1% (n = 131 clusters), which is slightly lower than the success rate for single-residue glycans on all-quality experimental data of 46.1% (n = 167). The modified residues dataset is filtered similarly to our other polymer test sets: it contains only modified residues in polymer chains with low homology to the training set (Methods). See Extended Data Table 1 for detailed results, and Extended Data Fig. 6 for examples of predicted protein, DNA and RNA structures with covalent modifications, including analysis of the impact of phosphorylation on predictions.

While expanding in modelling abilities, AF3 has also improved in protein complex accuracy relative to AlphaFold-Multimer (v.2.3)[7,8]. Generally, protein–protein prediction success (DockQ > 0.23)[40] has increased (paired Wilcoxon signed-rank test, $P = 1.8 \times 10^{-18}$), with antibody–protein interaction prediction in particular showing a marked improvement (Fig. 1c (right); paired Wilcoxon signed-rank test, $P = 6.5 \times 10^{-5}$, predictions top-ranked from 1,000 rather than the typical 5 seeds; further details are provided in Fig. 5a). Protein monomer LDDT improvement is also significant (paired Wilcoxon signed-rank test, $P = 1.7 \times 10^{-34}$). AF3 has a very similar dependence on MSA depth to AlphaFold-Multimer v.2.3; proteins with shallow MSAs are predicted with lower accuracy (a comparison of the dependence of single-chain LDDT on MSA depth is shown in Extended Data Fig. 7a).

Predicted confidences track accuracy

As with AF2, AF3 confidence measures are well calibrated with accuracy. Our confidence analysis is performed on the recent PDB evaluation set, with no homology filtering and including peptides. The ligands category is filtered to high-quality experimental structures as described above, and considers standard non-bonded ligands only. See Extended Data Fig. 8 for a similar assessment on bonded ligand and other interfaces. All statistics are cluster-weighted (Methods) and consider the top-ranked prediction only (ranking details are provided in Supplementary Methods 5.9.3).

In Fig. 4a (top row), we plot the chain pair interface-predicted TM (ipTM) score[41] (Supplementary Methods 5.9.1) against interface accuracy measures: protein–protein DockQ, protein–nucleic interface LDDT (iLDDT) and protein–ligand success, with success defined as the percentage of examples under thresholded pocket-aligned r.m.s.d. values. In Fig. 4a (bottom row), we plot the average pLDDT per protein, nucleotide or ligand entity against our bespoke LDDT_to_polymer metric (metrics details are provided in the Methods), which is closely related to the training target of the pLDDT predictor.

In Fig. 4b–e, we highlight a single example prediction of 7T82, in which per-atom pLDDT colouring identifies unconfident chain tails, somewhat confident interfaces and otherwise confident secondary structure. In Fig. 4c, the same prediction is coloured by chain, along with DockQ interface scores in Fig. 4d and per-chain colouring displayed on the axes for reference. We see from Fig. 4e that PAE confidence is high for pink–grey and blue–orange residue pairs for which DockQ > 0.7, and least confident about pink–orange and pink–blue residue pairs that have DockQ ≈ 0. A similar PAE analysis of an example with protein and nucleic acid chains is shown in Extended Data Fig. 5c,d.

Model limitations

We note model limitations of AF3 with respect to stereochemistry, hallucinations, dynamics and accuracy for certain targets.

On stereochemistry, we note two main classes of violations. The first is that the model outputs do not always respect chirality (Fig. 5b), despite the model receiving reference structures with correct chirality as input features. To address this in the PoseBusters benchmark, we included a penalty for chirality violation in our ranking formula for model predictions. Despite this, we still observe a chirality violation rate of 4.4% in the benchmark. The second class of stereochemical violations is a tendency of the model to occasionally produce overlapping (clashing) atoms in the predictions. This sometimes manifests as extreme violations in homomers in which entire chains have been observed to overlap (Fig. 5e). Penalizing clashes during ranking (Supplementary Methods 5.9.3) reduces the occurrence of this failure mode but does not eliminate them. Almost all remaining clashes occur for protein–nucleic complexes with both greater than 100 nucleotides and greater than 2,000 residues in total.

We note that the switch from the non-generative AF2 model to the diffusion-based AF3 model introduces the challenge of spurious structural order (hallucinations) in disordered regions (Fig. 5d and Extended Data Fig. 1). Although hallucinated regions are typically marked as very low confidence, they can lack the distinctive ribbon-like appearance

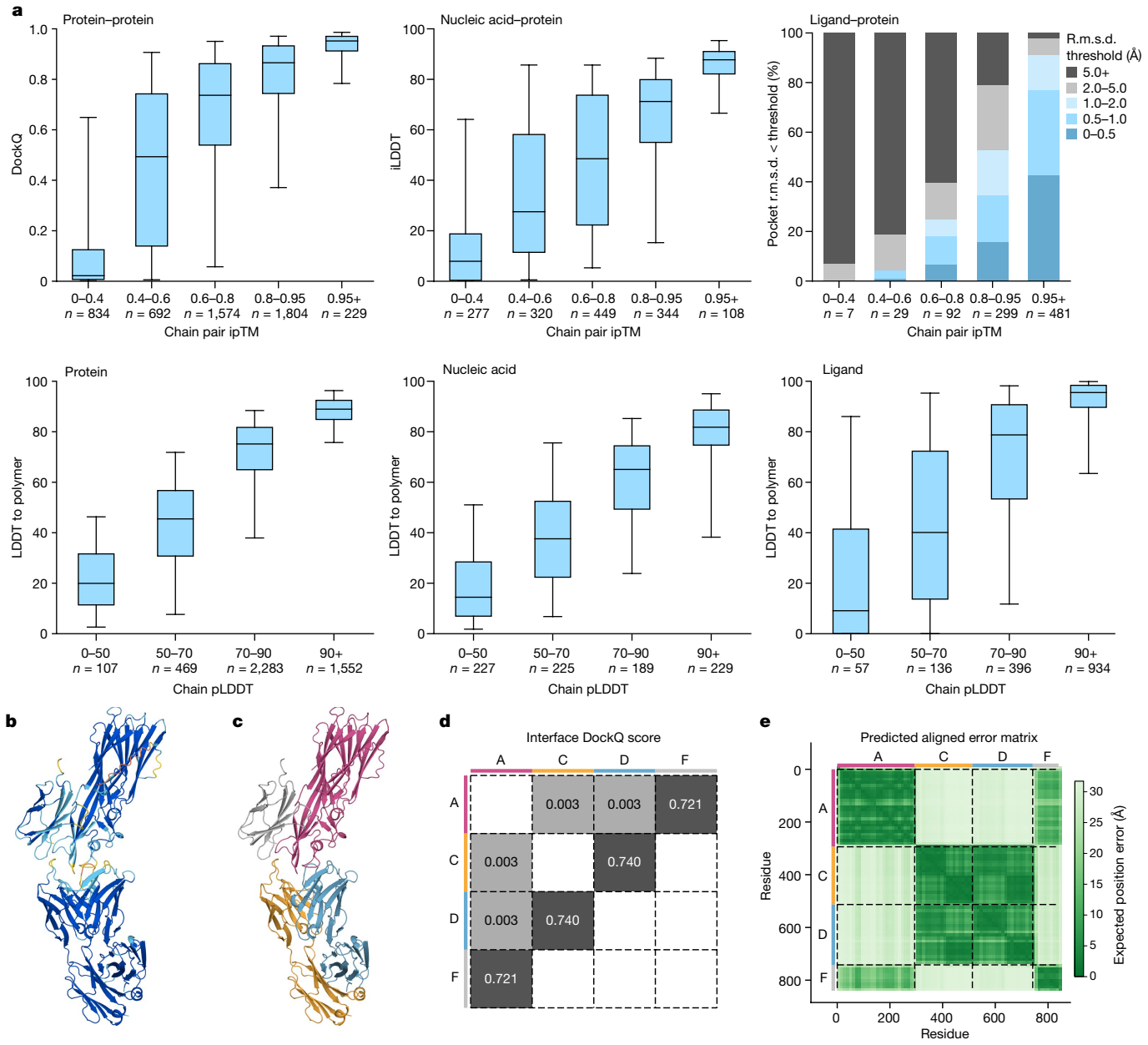


Fig. 4 | AF3 confidences track accuracy. a, The accuracy of protein-containing interfaces as a function of chain pair ipTM (top). Bottom, the LDDT-to-polymer accuracy was evaluated for various chain types as a function of chain-averaged pLDDT. The box plots show the 25–75% confidence intervals (box limits), the median (centre line) and the 5–95% confidence intervals (whiskers). *n* values report the number of clusters in each band. **b**, The predicted structure of PDB 7T82 coloured by pLDDT (orange, 0–50; yellow, 50–70; cyan, 70–90; and blue, 90–100). **c**, The same prediction coloured by chain. **d**, DockQ scores for protein–protein interfaces. **e**, PAE matrix of same prediction (darker is more confident), with chain colouring of **c** on the side bars. The dashed black lines indicate the chain boundaries.

that AF2 produces in disordered regions. To encourage ribbon-like predictions in AF3, we use distillation training from AF2 predictions, and we add a ranking term to encourage results with more solvent accessible surface area[36].

A key limitation of protein structure prediction models is that they typically predict static structures as seen in the PDB, not the dynamical behaviour of biomolecular systems in solution. This limitation persists for AF3, in which multiple random seeds for either the diffusion head or the overall network do not produce an approximation of the solution ensemble.

In some cases, the modelled conformational state may not be correct or comprehensive given the specified ligands and other inputs. For example, E3 ubiquitin ligases natively adopt an open conformation in an apo state and have been observed only in a closed state when bound to ligands, but AF3 exclusively predicts the closed state for both holo and apo systems[42] (Fig. 5c). Many methods have been developed, particularly around MSA resampling, that assist in generating diversity from previous AlphaFold models[43–45] and may also assist in multistate prediction with AF3.

Despite the large advance in modelling accuracy in AF3, there are still many targets for which accurate modelling can be challenging. To obtain the highest accuracy, it may be necessary to generate a large number of predictions and rank them, which incurs an extra computational cost. A class of targets in which we observe this effect

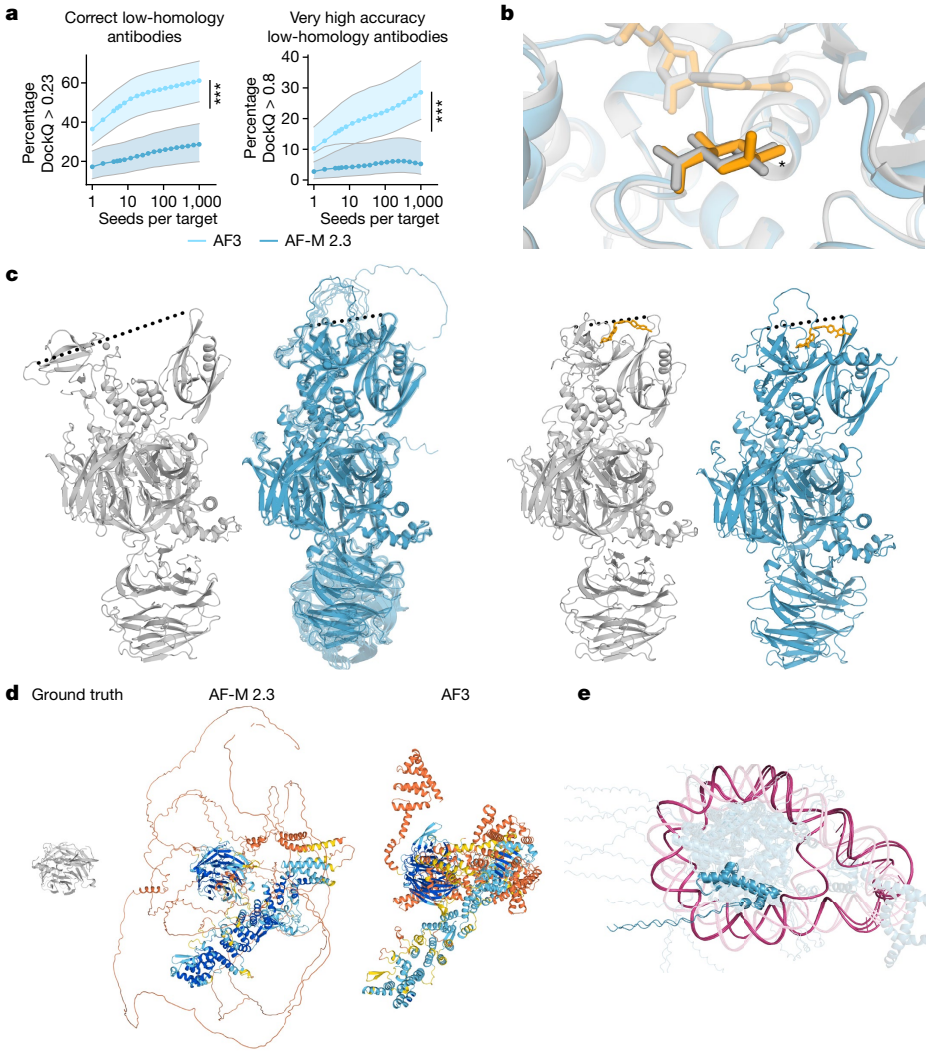


Fig. 5 | Model limitations. a, Antibody prediction quality increases with the number of model seeds. The quality of top-ranked, low-homology antibody–antigen interface predictions as a function of the number of seeds. Each datapoint shows the mean over 1,000 random samples (with replacement) of seeds to rank over, out of 1,200 seeds. Confidence intervals are 95% bootstraps over 10,000 resamples of cluster scores at each datapoint. Samples per interface are ranked by protein–protein ipTM. Significance tests were performed using by two-sided Wilcoxon signed-rank tests. $n = 65$ clusters. Exact P values were as follows: 2.0×10^{-5} (percentage correct) and $P = 0.009$ (percentage very high accuracy). **b**, Prediction (coloured) and ground-truth (grey) structures of *Thermotoga maritima* α-glucuronidase and beta-D-glucuronic acid—a target from the PoseBusters set (PDB: 7CTM). AF3 predicts alpha-D-glucuronic acid; the differing chiral centre is indicated by an asterisk. The prediction shown is top-ranked by ligand–protein ipTM and with a chirality and clash penalty. **c**, Conformation coverage is limited. Ground-truth structures (grey) of cereblon in open (apo, PDB: 8CVP; left) and closed (holo mezigdomide-bound, PDB: 8D7U; right) conformations. Predictions (blue) of both apo (with 10 overlaid samples) and holo structures are in the closed conformation. The dashed lines indicate the distance between the N-terminal Lon protease-like and C-terminal thalidomide-binding domain. **d**, A nuclear pore complex with 1,854 unresolved residues (PDB: 7F60). The ground truth (left) and predictions from AlphaFold-Multimer v.2.3 (middle) and AF3 (right) are shown. **e**, Prediction of a trinucleosome with overlapping DNA (pink) and protein (blue) chains (PDB: 7PEU); highlighted are overlapping protein chains B and J and self-overlapping DNA chain AA. Unless otherwise stated, predictions are top-ranked by our global complex ranking metric with chiral mismatch and steric clash penalties (Supplementary Methods 5.9.1).

strongly is antibody–antigen complexes, similar to other recent work[46]. Figure 5a shows that, for AF3, top-ranked predictions keep improving with more model seeds, even at as many as 1,000 (Wilcoxon signed-rank test between 5 and 1,000 seeds, $P = 2.0 \times 10^{-5}$ for percentage correct and $P = 0.009$ for percentage very high accuracy; ranking by protein–protein interface ipTM). This large improvement with many seeds is not observed in general for other classes of molecules (Extended Data Fig. 7b). Using only one diffusion sample per model seed for the AF3 predictions rather than five (not illustrated) does not change the results significantly, indicating that running more model seeds is necessary for antibody score improvements, rather than just more diffusion samples.

Discussion

The core challenge of molecular biology is to understand and ultimately regulate the complex atomic interactions of biological systems. The AF3 model takes a large step in this direction, demonstrating that it is possible to accurately predict the structure of a wide range of biomolecular systems in a unified framework. Although there are still substantial challenges to achieve highly accurate predictions across all interaction types, we demonstrate that it is possible to build a deep-learning system that shows strong coverage and generalization for all of these interactions. We also demonstrate that the lack of cross-entity evolutionary information is not a substantial blocker to progress in predicting

these interactions and, moreover, substantial improvement in antibody results suggests AlphaFold-derived methods are able to model the chemistry and physics of classes of molecular interactions without dependence on MSAs. Finally, the large improvement in protein–ligand structure prediction shows that it is possible to handle the wide diversity of chemical space within a general deep-learning framework and without resorting to an artificial separation between protein structure prediction and ligand docking.

The development of bottom-up modelling of cellular components is a key step in unravelling the complexity of molecular regulation within the cell, and the performance of AF3 shows that developing the right deep-learning frameworks can massively reduce the amount of data required to obtain biologically relevant performance on these tasks and amplify the impact of the data already collected. We expect that structural modelling will continue to improve not only due to advances in deep learning but also because continuing methodological advances in experimental structure determination, such as the substantial improvements in cryo-electron microscopy and tomography, will provide a wealth of new training data to further the improve the generalization ability of such models. The parallel developments of experimental and computational methods promise to propel us further into an era of structurally informed biological understanding and therapeutic development.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07487-w.

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Kreitz, J. et al. Programmable protein delivery with a bacterial contractile injection system. *Nature* **616**, 357–364 (2023).
3. Lim, Y. et al. In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science* **381**, eadi3448 (2023).
4. Mosalaganti, S. et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).
5. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.15019 (2022).
6. Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target. Ther.* **8**, 115 (2023).
7. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2022).
8. Žídek, A. AlphaFold v.2.3.0 Technical Note. *GitHub* https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md (2022).
9. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Opin. Struct. Biol.* **79**, 102548 (2023).
10. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* https://doi.org/10.1126/science.abj8754 (2021).
12. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.21.500999 (2022).
13. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
14. Moriwaki, Y. Post on X. *X* https://x.com/Ag_smith/status/1417063635000598528?lang=en-GB (2021).
15. Baek, M. Post on X. *X* https://x.com/minkbaek/status/1417538291709071362?lang=en (2021).
16. Qiao, Z. et al. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **6**, 195–208 (2024).
17. Nakata, S., Mori, Y. & Tanaka, S. End-to-end protein–ligand complex structure generation with diffusion-based generative models. *BMC Bioinform.* **24**, 233 (2023).
18. Baek, M. et al. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**, 117–121 (2024).
19. Townshend, R. J. L. et al. Geometric deep learning of RNA structure. *Science* **373**, 1047–1051 (2021).
20. Jiang, D. et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions. *J. Med. Chem.* **64**, 18209–18232 (2021).
21. Jiang, H. et al. Predicting protein–ligand docking structure with graph neural network. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.2c00127 (2022).
22. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: diffusion steps, twists, and turns for molecular docking. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.01776 (2022).
23. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. & Jaakkola, T. EquiBind: Geometric deep learning for drug binding structure prediction. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 20503–20521 (PMLR, 2022).
24. Liao, Z. et al. DeepDock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information. In *Proc. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 311–317 (IEEE, 2019).
25. Lu, W. et al. TANKBind: trigonometry-aware neural networks for drug-protein binding structure prediction. *Adv. Neural Inf. Process. Syst.* **35**, 7236–7249 (2022).
26. Zhou, G. et al. Uni-Mol: a universal 3D molecular representation learning framework. Preprint at *ChemRxiv* https://chemrxiv.org/engage/chemrxiv/article-details/6402990d37e01856dc1d1581 (2023).
27. Shen, T. et al. E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. Preprint at *arXiv* https://arxiv.org/abs/2207.01586 (2022).
28. van Dijk, M. & Bonvin, A. M. J. J. Pushing the limits of what is achievable in protein–DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res.* **38**, 5634–5647 (2010).
29. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
30. Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).
31. Das, R. et al. Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins* **91**, 1747–1770 (2023).
32. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
33. Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **35**, 26565–26577 (2022).
34. Wang, Y., Elhag, A. A., Jaitly, N., Susskind, J. M. & Bautista, M. A. Generating molecular conformer fields. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2311.17932 (2023).
35. Ji, Z., et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248 (2023).
36. Del Conte, A. et al. Critical assessment of protein intrinsic disorder prediction (CAID)—results of round 2. *Proteins* **91**, 1925–1934 (2023).
37. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
38. Miller, E. B. et al. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.1c00136 (2021).
39. Chen, K., Zhou, Y., Wang, S. & Xiong, P. RNA tertiary structure modeling with BRiQ potential in CASP15. *Proteins* **91**, 1771–1778 (2023).
40. Basu, S. & Wallner, B. DockQ: a quality measure for protein-protein docking models. *PLoS ONE* **11**, e0161879 (2016).
41. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
42. Watson, E. R. et al. Molecular glue CELMoD compounds are regulators of cereblon conformation. *Science* **378**, 549–553 (2022).
43. Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
44. del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* https://doi.org/10.7554/eLife.75751 (2022).
45. Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885 (2022).
46. Wallner, B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).
47. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
48. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024, corrected publication 2024

Methods

Full algorithm details

Extensive explanations of the components are available in Supplementary Methods 2–5. Moreover, pseudocode is available in Supplementary Algorithms 1–31, network diagrams in Figs. 1d and 2a–c and Supplementary Fig. 2, input features in Supplementary Table 5 and additional hyperparameters for training in Supplementary Tables 3, 4 and 7.

Training regime

No structural data used during training were released after 30 September 2021 and, for the model used in PoseBusters evaluations, we filtered out PDB[32] structures released after 30 September 2021. One optimizer step uses a mini batch of 256 input data samples and during initial training 256 × 48 = 12,288 diffusion samples. For fine-tuning, the number of diffusion samples is reduced to 256 × 32 = 8,192. The model is trained in three stages—the initial training with a crop size of 384 tokens and two sequential fine tuning stages with crop sizes of 640 and 768 tokens. Further details are provided in Supplementary Methods 5.2.

Inference regime

No inference time templates or reference ligand position features were released after 30 September 2021, and in the case of PoseBusters evaluation, an earlier cut-off date of 30 September 2019 was used. The model can be run with different random seeds to generate alternative results, with a batch of diffusion samples per seed. Unless otherwise stated, all results are generated by selecting the top confidence sample from running 5 seeds of the same trained model, with 5 diffusion samples per model seed, for a total of 25 samples to choose from. Standard crystallization aids are excluded from predictions (Supplementary Table 8).

Results are shown for the top-ranked sample and sample ranking depends on whether trying to select the overall best output globally, or the best output for some chain, interface or modified residue. Global ranking uses a mix of pTM and ipTM along with terms to reduce cases with large numbers of clashes and increase rates of disorder; individual chain ranking uses a chain specific pTM measure; interface ranking uses a bespoke ipTM measure for the relevant chain pair; and modified residue ranking uses average pLDDT over the residue of interest (Supplementary Methods 5.9.3).

Metrics

Evaluation compares a predicted structure to the corresponding ground-truth structure. If the complex contains multiple identical entities, assignment of the predicted units to the ground-truth units is found by maximizing LDDT. Assignment in local symmetry groups of atoms in ligands is solved by exhaustive search over the first 1,000 per-residue symmetries as given by RDKit.

We measure the quality of the predictions with DockQ, LDDT or pocket-aligned r.m.s.d. For nucleic–protein interfaces, we measure interface accuracy through iLDDT, which is calculated from distances between atoms across different chains in the interface. DockQ and iLDDT are highly correlated (Extended Data Fig. 9), so the standard cut-offs for DockQ can be translated to equivalent iLDDT cut-offs. Nucleic acid LDDTs (intrachains and interface) were calculated with an inclusion radius of 30 Å compared with the usual 15 Å used for proteins, owing to their larger scale. For confidence calibration assessment, we use a bespoke LDDT (LDDT_to_polymer) metric that considers differences from each atom of a given entity to any C^{α} or C1′ polymer atom within its inclusion radius. This is closely related to how the confidence prediction is trained (Supplementary Methods 4.3.1).

Pocket-aligned r.m.s.d. is computed as follows: the pocket is defined as all heavy atoms within 10 Å of any heavy atom of the ligand, restricted to the primary polymer chain for the ligand or modified residue being scored, and further restricted to only backbone atoms for proteins. The primary polymer chain is defined variously: for PoseBusters, it is the protein chain with the most atoms within 10 Å of the ligand; for bonded ligand scores, it is the bonded polymer chain; and for modified residues, it is the chain in which the residue is contained (minus that residue). The pocket is used to align the predicted structure to the ground-truth structure with least-squares rigid alignment and then the r.m.s.d. is computed on all heavy atoms of the ligand.

Recent PDB evaluation set

General model evaluation was performed on our recent PDB set consisting of 8,856 PDB complexes released between 1 May 2022 and 12 January 2023. The set contains almost all PDB complexes released during that period that are less than 5,120 model tokens in size (Supplementary Methods 6.1). Single chains and interfaces within each structure were scored separately rather than only looking at full complex scores, and clustering was then applied to chains and interfaces so that scores could be aggregated first within clusters and then across clusters for mean scores, or using a weighting of inverse cluster size for distributional statistics (Supplementary Methods 6.2 and 6.4).

Evaluation on ligands excludes standard crystallization aids (Supplementary Table 8), our ligand exclusion list (Supplementary Table 9) and glycans (Supplementary Table 10). Bonded and non-bonded ligands are evaluated separately. Ions are only included when specifically mentioned (Supplementary Table 11).

The recent PDB set is filtered to a low homology subset (Supplementary Methods 6.1) for some results where stated. Homology is defined as sequence identity to sequences in the training set and is measured by template search (Supplementary Methods 2.4). Individual polymer chains in evaluation complexes are filtered out if the maximum sequence identity to chains in the training set is greater than 40%, where sequence identity is the percentage of residues in the evaluation set chain that are identical to the training set chain. Individual peptide chains (protein chains with less than 16 residues) are always filtered out. For polymer–polymer interfaces, if both polymers have greater than 40% sequence identity to two chains in the same complex in the training set, then the interface is filtered out. For interfaces to a peptide, the interface is filtered out if the non-peptide entity has greater than 40% sequence identity to any chain in the training set.

To compare the quality of prediction of protein–protein interfaces and protein monomers against that of AlphaFold-Multimer (v.2.3)[8], and to compare the dependence of single-protein-chain prediction quality on MSA depth, we restrict the low-homology recent PDB set to complexes with fewer than 20 protein chains and fewer than 2,560 tokens. We compare against unrelaxed AlphaFold-Multimer v.2.3 predictions.

To study antibody-antigen interface prediction, we filter the low homology recent PDB set to complexes that contain at least one protein–protein interface where one of the protein chains is in one of the two largest PDB chain clusters (these clusters are representative of antibodies). We further filter to complexes with at most 2,560 tokens and with no unknown amino acids in the PDB to allow extensive comparison against relaxed predictions of AlphaFold-Multimer v2.3. That leaves 71 antibody–antigen complexes, containing 166 antibody–antigen interfaces spanning 65 interface clusters.

MSA depth analysis (Extended Data Fig. 7a) was based on computing the normalized number of effective sequences (N_{eff}) for each position of a query sequence. Per-residue N_{eff} values were obtained by counting the number of non-gap residues in the MSA for this position and weighting the sequences using the N_{eff} scheme[49] with a threshold of 80% sequence identity measured on the region that is non-gap in either sequence.

Nucleic acid prediction baseline

For benchmarking performance on nucleic acid structure prediction, we report baseline comparisons to an existing machine learning system for protein–nucleic acid and RNA tertiary structure prediction, RoseTTAFold2NA[18]. We run the open source RF2NA[50] with the same MSAs as those that were used for AF3 predictions. For comparison

between AF3 and RF2NA, a subset of our recent PDB set was chosen to meet the RF2NA criteria (<1,000 total residues and nucleotides). As RF2NA was not trained to predict systems with DNA and RNA, analysis is limited to targets with only one nucleic acid type. No system was publicly available at time of writing for baseline comparisons on data with arbitrary combinations of biomolecular types in PDB.

As an additional baseline for RNA tertiary structure prediction, we evaluate AF3 performance on CASP15 RNA targets that were publicly available as of 1 December 2023 (R1116/8S95, R1117/8FZA, R1126 (downloaded from the CASP15 website https://predictioncenter.org/casp15/TARGETS_PDB/R1126.pdb), R1128/8BTZ, R1136/7ZJ4, R1138/[7PTK/7PTL], R1189/7YR7 and R1190/7YR6). We compare the top-1 ranked predictions and, where multiple ground-truth structures exist (R1136), the prediction is scored against the closest state. We display comparisons to RF2NA as a representative machine learning system; AIchemy_RNA2 as the top performing entrant with human intervention; and AIchemy_RNA as the top performing machine learning system. All entrants' predictions were downloaded from the CASP website and scored internally.

PoseBusters

While other analyses used an AlphaFold model trained on PDB data released before a cut-off of 30 September 2021, our PoseBusters analysis was conducted on a model (with identical architecture and similar training schedule) differing only in the use of an earlier 30 September 2019 cut-off. This analysis therefore did not include training data, inference time templates or 'ref_pos' features released after this date.

Inference was performed on the asymmetric unit from specified PDBs, with the following minor modifications. In several PDB files, chains clashing with the ligand of interest were removed (7O1T, 7PUV, 7SCW, 7WJB, 7ZXV, 8AIE). Another PDB entry (8F4J) was too large to inference the entire system (over 5,120 tokens), so we included only protein chains within 20 Å of the ligand of interest. Five model seeds, each with five diffusion samples, were produced per target, resulting in 25 predictions, which were ranked by quality and predicted accuracy: the ranking score was calculated from an ipTM aggregate (Supplementary Methods 5.9.3 (point 3)), then further divided by 100 if the ligand had chirality errors or had clashes with the protein.

For pocket-aligned r.m.s.d., first alignment between the predicted and ground-truth structures was conducted by aligning to the ground-truth pocket backbone atoms (CA, C or N atoms within 10 Å of the ligand of interest) from the primary protein chain (the chain with the greatest number of contacts within 10 Å of the ligand). The PoseBusters Python package v.0.2.7[51] was used to score r.m.s.d. and violations from the pocket-aligned predictions.

While AlphaFold models are 'blind' to the protein pocket, docking is often performed with knowledge of the protein pocket residues. For example, Uni-Mol specifies the pocket as any residue within 6 Å of the heavy atoms in the ligand of interest[26]. To evaluate the ability of AF3 to dock ligands accurately when given pocket information, we fine-tuned a 30 September 2019 cut-off AF3 model with an additional token feature specifying pocket–ligand pairs (Supplementary Methods 2.8). Specifically, an additional token feature was introduced, set to true for a ligand entity of interest and any pocket residues with heavy atoms within 6 Å of the ligand entity. At training time, a single random ligand entity is chosen to use in this feature. Note that multiple ligand chains with the same entity (CCD code) may be selected. At inference time, the ligand entity was chosen based on the ligand of interest's CCD code, so again multiple ligand chains were occasionally chosen. The results of this analysis are shown in Extended Data Fig. 4.

Model performance analysis and visualization

Data analysis used Python v.3.11.7 (https://www.python.org/), NumPy v.1.26.3 (https://github.com/numpy/numpy), SciPy v.1.9.3 (https://www.scipy.org/), seaborn v.0.12.2 (https://github.com/mwaskom/seaborn), Matplotlib v.3.6.1 (https://github.com/matplotlib/matplotlib), pandas v.2.0.3 (https://github.com/pandas-dev/pandas), statsmodels v.0.12.2 (https://github.com/statsmodels/statsmodels), RDKit v.4.3.0 (https://github.com/rdkit/rdkit) and Colab (https://research.google.com/colaboratory). TM-align v.20190822 (https://zhanglab.dcmb.med.umich.edu/TM-align/) was used for computing TM-scores. Structure visualizations were created in Pymol v.2.55.5 (https://github.com/schrodinger/pymol-open-source).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All scientific datasets used to create training and evaluation inputs are freely available from public sources. Structures from the PDB were used for training and as templates (https://files.wwpdb.org/pub/pdb/data/assemblies/mmCIF/; sequence clusters are available at https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt; sequence data are available at https://files.wwpdb.org/pub/pdb/derived_data/). Training used a version of the PDB downloaded 12 January 2023, while template search used a version downloaded 28 September 2022. We also used the Chemical Components Dictionary downloaded on 19 October 2023 (https://www.wwpdb.org/data/ccd). We show experimental structures from the PDB under accession numbers 7PZB (ref. 52), 7PNM (ref. 53), 7TQL (ref. 54), 7AU2 (ref. 55), 7U8C (ref. 56), 7URD (ref. 57), 7WUX (ref. 58), 7QIE (ref. 59), 7T82 (ref. 60), 7CTM (ref. 61), 8CVP (ref. 42), 8D7U (ref. 42), 7F60 (ref. 62), 8BTI (ref. 63), 7KZ9 (ref. 64), 7XFA (ref. 65), 7PEU (ref. 66), 7SDW (ref. 67), 7TNZ (ref. 68), 7R6R (ref. 69), 7USR (ref. 70) and 7Z1K (ref. 71). We also used the following publicly available databases for training or evaluation. Detailed usage is described in Supplementary Methods 2.2 and 2.5.2. UniRef90 v.2020_01 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/), UniRef90 v.2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/), UniRef90 v.2022_05 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_05/uniref/), Uniclust30 v.2018_08 (https://wwwuser.gwdg.de/~compbiol/uniclust/2018_08/), Uniclust30 v.2021_03 (https://wwwuser.gwdg.de/~compbiol/uniclust/2021_03/), MGnify clusters v.2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/), MGnify clusters v.2022_05 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2022_05/), BFD (https://bfd.mmseqs.com), RFam v.14.9 (https://ftp.ebi.ac.uk/pub/databases/Rfam/14.9/), RNAcentral v.21.0 (https://ftp.ebi.ac.uk/pub/databases/RNAcentral/releases/21.0/), Nucleotide Database (as of 23 February 2023) (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz), JASPAR 2022 (https://jaspar.elixir.no/downloads/; see https://jaspar.elixir.no/profile-versions for version information), SELEX protein sequences from the supplementary tables of ref. 72 and SELEX protein sequences from the supplementary tables of ref. 73.

Code availability

AlphaFold 3 will be available as a non-commercial usage only server at https://www.alphafoldserver.com, with restrictions on allowed ligands and covalent modifications. Pseudocode describing the algorithms is available in the Supplementary Information. Code is not provided.

49. Wu, T., Hou, J., Adhikari, B. & Cheng, J. Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics* **36**, 1091–1098 (2020).
50. DiMaio, F. RF2NA v.0.2. *GitHub* https://github.com/uw-ipd/RoseTTAFold2NA/releases/tag/v0.2 (2023).
51. Buttenschoen, M. PoseBusters v.0.2.7. *GitHub* https://github.com/maabuu/posebusters/releases/tag/v0.2.7 (2023).

52. Werel, L. et al. Structural basis of dual specificity of *Sinorhizobium meliloti* Clr, a cAMP and cGMP receptor protein. *MBio* **14**, e0302822 (2023).
53. Wang, C. et al. Antigenic structure of the human coronavirus OC43 spike reveals exposed and occluded neutralizing epitopes. *Nat. Commun.* **13**, 2921 (2022).
54. Lapointe, C. P. et al. eIF5B and eIF1A reorient initiator tRNA to allow ribosomal subunit joining. *Nature* **607**, 185–190 (2022).
55. Wilson, L. F. L. et al. The structure of EXTL3 helps to explain the different roles of bi-domain exostosins in heparan sulfate synthesis. *Nat. Commun.* **13**, 3314 (2022).
56. Liu, X. et al. Highly active CAR T cells that bind to a juxtamembrane region of mesothelin and are not blocked by shed mesothelin. *Proc. Natl Acad. Sci. USA* **119**, e2202439119 (2022).
57. Liu, Y. et al. Mechanisms and inhibition of Porcupine-mediated Wnt acylation. *Nature* **607**, 816–822 (2022).
58. Kurosawa, S. et al. Molecular basis for enzymatic aziridine formation via sulfate elimination. *J. Am. Chem. Soc.* **144**, 16164–16170 (2022).
59. Boffey, H. K. et al. Development of selective phosphatidylinositol 5-phosphate 4-kinase γ inhibitors with a non-ATP-competitive, allosteric binding mode. *J. Med. Chem.* **65**, 3359–3370 (2022).
60. Buckley, P. T. et al. Multivalent human antibody-centyrin fusion protein to prevent and treat *Staphylococcus aureus* infections. *Cell Host Microbe* **31**, 751–765 (2023).
61. Mohapatra, S. B. & Manoj, N. Structural basis of catalysis and substrate recognition by the NAD(H)-dependent α-d-glucuronidase from the glycoside hydrolase family 4. *Biochem. J.* **478**, 943–959 (2021).
62. Gao, X. et al. Structural basis for *Sarbecovirus* ORF6 mediated blockage of nucleocytoplasmic transport. *Nat. Commun.* **13**, 4782 (2022).
63. Atkinson, B. N. et al. Designed switch from covalent to non-covalent inhibitors of carboxylesterase Notum activity. *Eur. J. Med. Chem.* **251**, 115132 (2023).
64. Luo, S. et al. Structural basis for a bacterial Pip system plant effector recognition protein. *Proc. Natl Acad. Sci. USA* **118**, e2019462118 (2021).
65. Liu, C. et al. Identification of monosaccharide derivatives as potent, selective, and orally bioavailable inhibitors of human and mouse galectin-3. *J. Med. Chem.* **65**, 11084–11099 (2022).
66. Dombrowski, M., Engeholm, M., Dienemann, C., Dodonova, S. & Cramer, P. Histone H1 binding to nucleosome arrays depends on linker DNA length and trajectory. *Nat. Struct. Mol. Biol.* **29**, 493–501 (2022).
67. Vecchioni, S. et al. Metal-mediated DNA nanotechnology in 3D: structural library by templated diffraction. *Adv. Mater.* **35**, e2210938 (2023).
68. Wang, W. & Pyle, A. M. The RIG-I receptor adopts two different conformations for distinguishing host from viral RNA ligands. *Mol. Cell* **82**, 4131–4144 (2022).
69. McGinnis, R. J. et al. A monomeric mycobacteriophage immunity repressor utilizes two domains to recognize an asymmetric DNA sequence. *Nat. Commun.* **13**, 4105 (2022).
70. Dietrich, M. H. et al. Nanobodies against Pfs230 block *Plasmodium falciparum* transmission. *Biochem. J.* **479**, 2529–2546 (2022).
71. Appel, L.-M. et al. The SPOC domain is a phosphoserine binding module that bridges transcription machinery with co- and post-transcriptional regulators. *Nat. Commun.* **14**, 166 (2023).
72. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
73. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

Acknowledgements We thank G. Arena, Ž. Avsec, A. Baryshnikov, R. Bates, M. Beck, A. Bond, N. Bradley-Schmieg, J. Cavojska, B. Coppin, E. Dupont, S. Eddy, M. Fiscato, R. Green, D. Hariharan, K. Holsheimer, N. Hurley, C. Jones, K. Kavukcuoglu, J. Kelly, E. Kim, A. Koivuniemi, O. Kovalevskiy, D. Lasecki, M. Last, A. Laydon, W. McCorkindale, S. Miller, A. Morris, L. Nicolaisen, E. Palmer, A. Paterson, S. Petersen, O. Purkiss, C. Shi, G. Thomas, G. Thornton and H. Tomlinson for their contributions.

Author contributions The equally contributing authors are alphabetically ordered, as are the remaining core contributor authors (excluding jointly supervising authors) and similar for all remaining non-supervising authors. D.H., M.J. and J.M.J. led the research. M.J., J.M.J. and P.K. developed research strategy. J. Abramson, V.B., T.G. and C.-C.H. led key research pillars. T.G. and A. Žídek led the technical framework for research. O.B., H.G. and S.S. coordinated and managed the research project. J. Abramson, J. Adler, E.A., A.J.B., J.B., V.B., A.I.C.-R., J.D., R.E., D.A.E., M.F., F.B.F., T.G., C.-C.H., M.J., J.M.J., Y.A.K., A. Potapenko, A. Pritzel, D.R., O.R., A.T., C.T., K.T., L.W., Z.W. and E.D.Z. developed the neural network architecture and training procedure. J. Abramson, A.J.B., J.B., V.B., C.B., S.W.B., A.B., A. Cherepanov, A.I.C.-R., A. Cowie, J.D., T.G., R.J., M.O., K.P., D.R., O.R., M.Z., A. Žemgulytė and A. Žídek developed the training, inference, data and evaluation infrastructure. J. Abramson, J. Adler, A.J.B., V.B., A.I.C.-R., R.E., D.A.E., T.G., D.H., M.J., J.M.J., P.K., K.P., A. Pritzel, O.R., P.S., S.S., A.S., K.T. and L.W. contributed to the writing of the paper. M.C., C.M.R.L. and S.Y. advised on the project.

Competing interests Author-affiliated entities have filed US provisional patent applications including 63/611,674, 63/611,638 and 63/546,444 relating to predicting 3D structures of molecule complexes using embedding neural networks and generative models. All of the authors other than A.B., Y.A.K. and E.D.Z. have commercial interests in the work described.

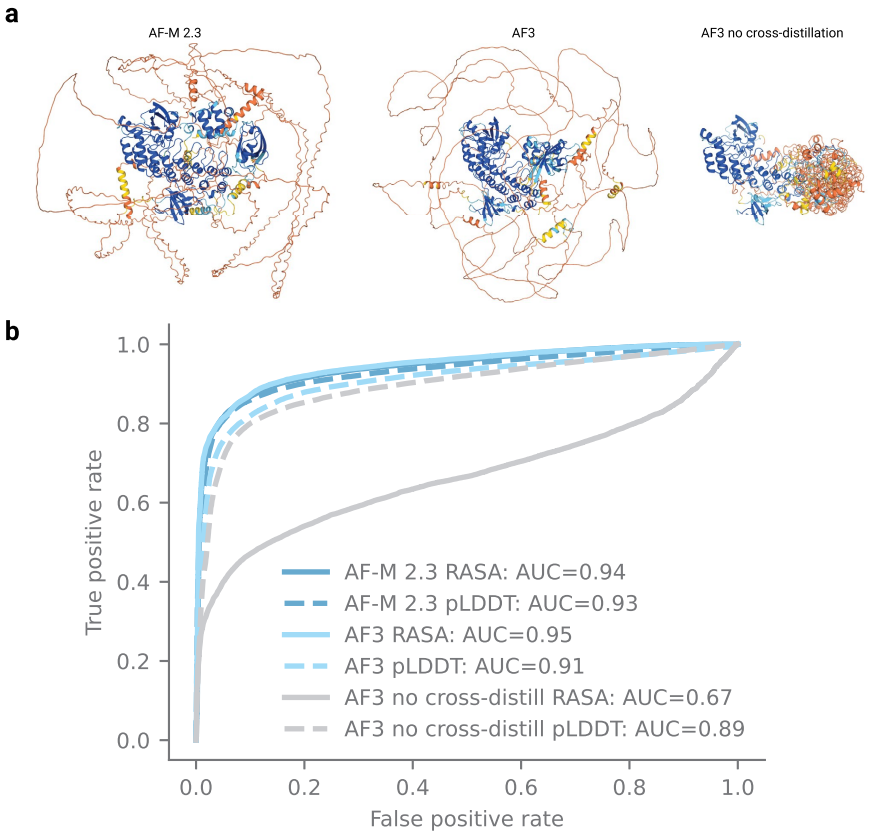
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07487-w.

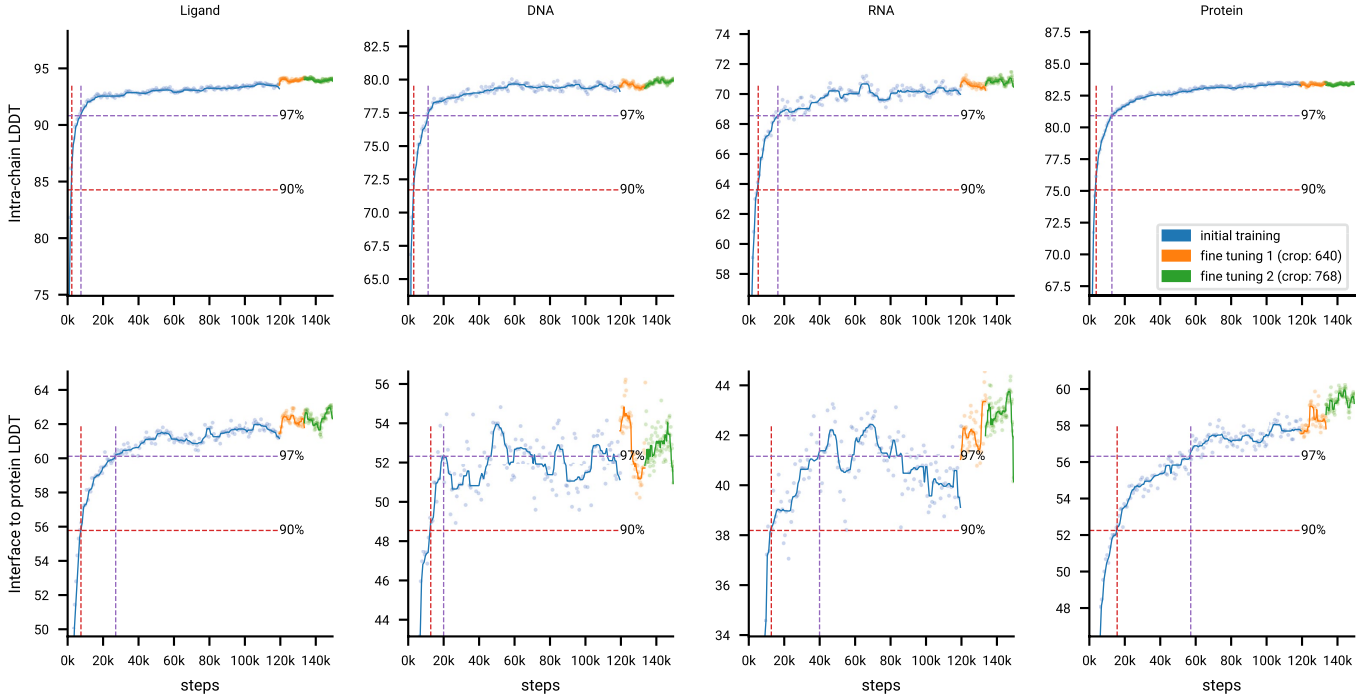
Correspondence and requests for materials should be addressed to Max Jaderberg, Demis Hassabis or John M. Jumper.

Peer review information *Nature* thanks Justas Dapkunas, Roland Dunbrack and Hashim Al-Hashimi for their contribution to the peer review of this work.

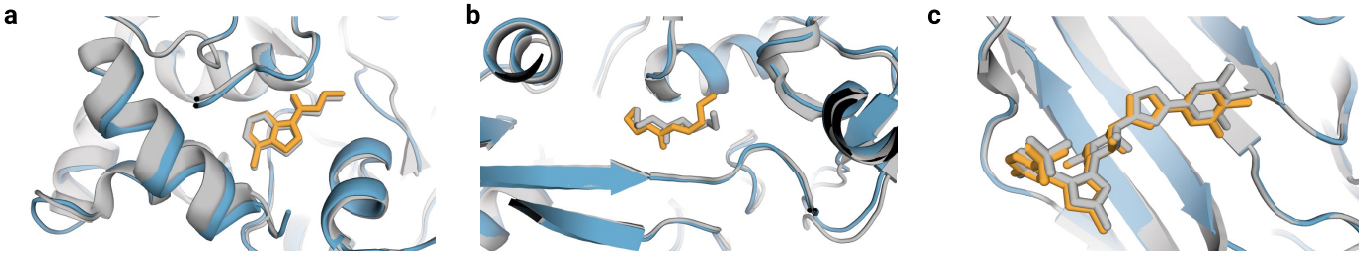
Reprints and permissions information is available at http://www.nature.com/reprints.



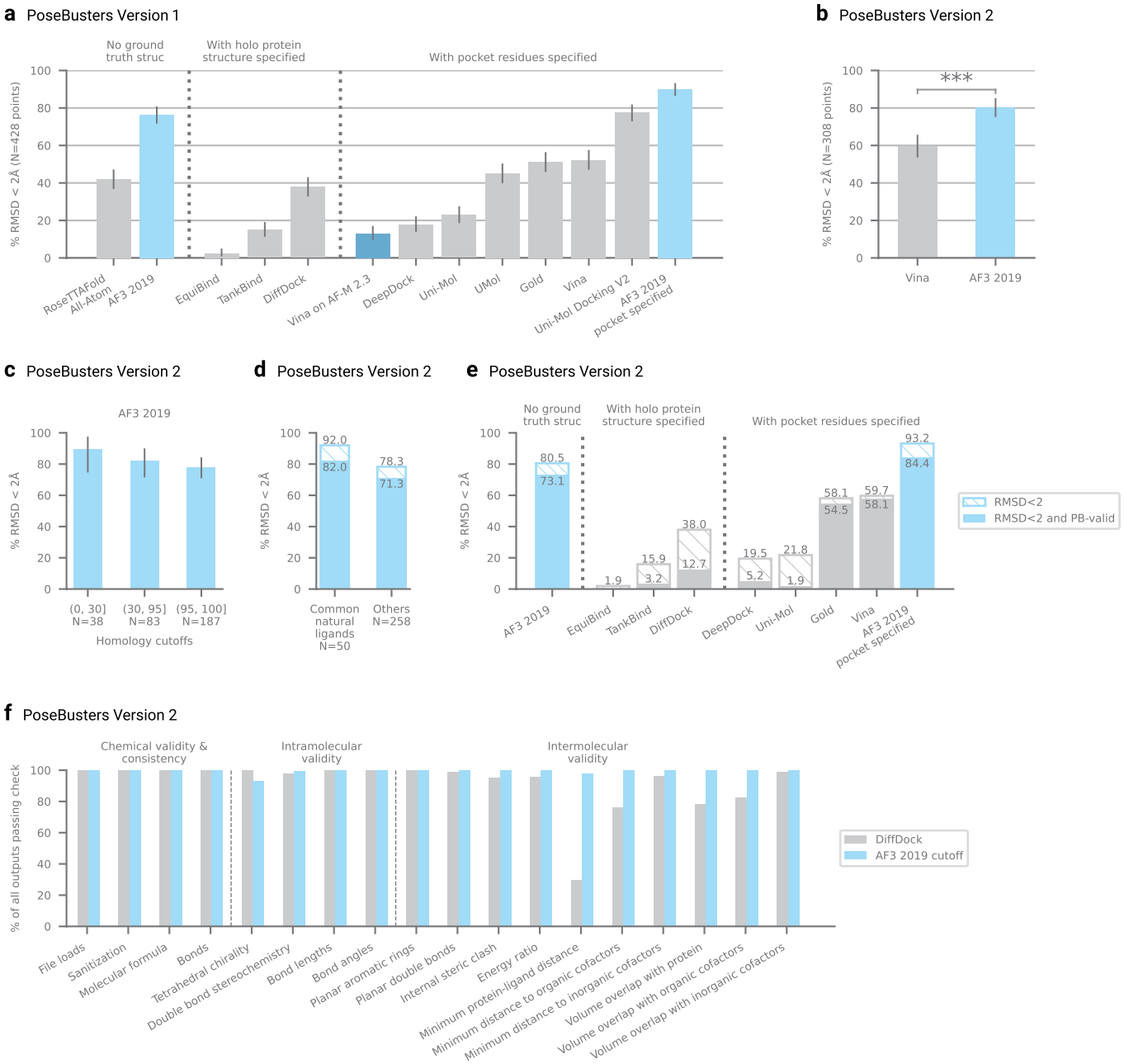
Extended Data Fig. 1 | Disordered region prediction. a, Example prediction for a disordered protein from AlphaFoldMultimer v2.3, AlphaFold 3, and AlphaFold 3 trained without the disordered protein PDB cross distillation set. Protein is DP02376 from the CAID 2 (Critical Assessment of protein Intrinsic Disorder prediction) set. Predictions coloured by pLDDT (orange: pLDDT <= 50, yellow: 50 < pLDDT <= 70, light blue: 70 < pLDDT <= 90, and dark blue: 90 <= pLDDT < 100). **b**, Predictions of disorder across residues in proteins in the CAID 2 set, which are also low homology to the AF3 training set. Prediction methods include RASA (relative accessible surface area) and pLDDT (N = 151 proteins; 46,093 residues).



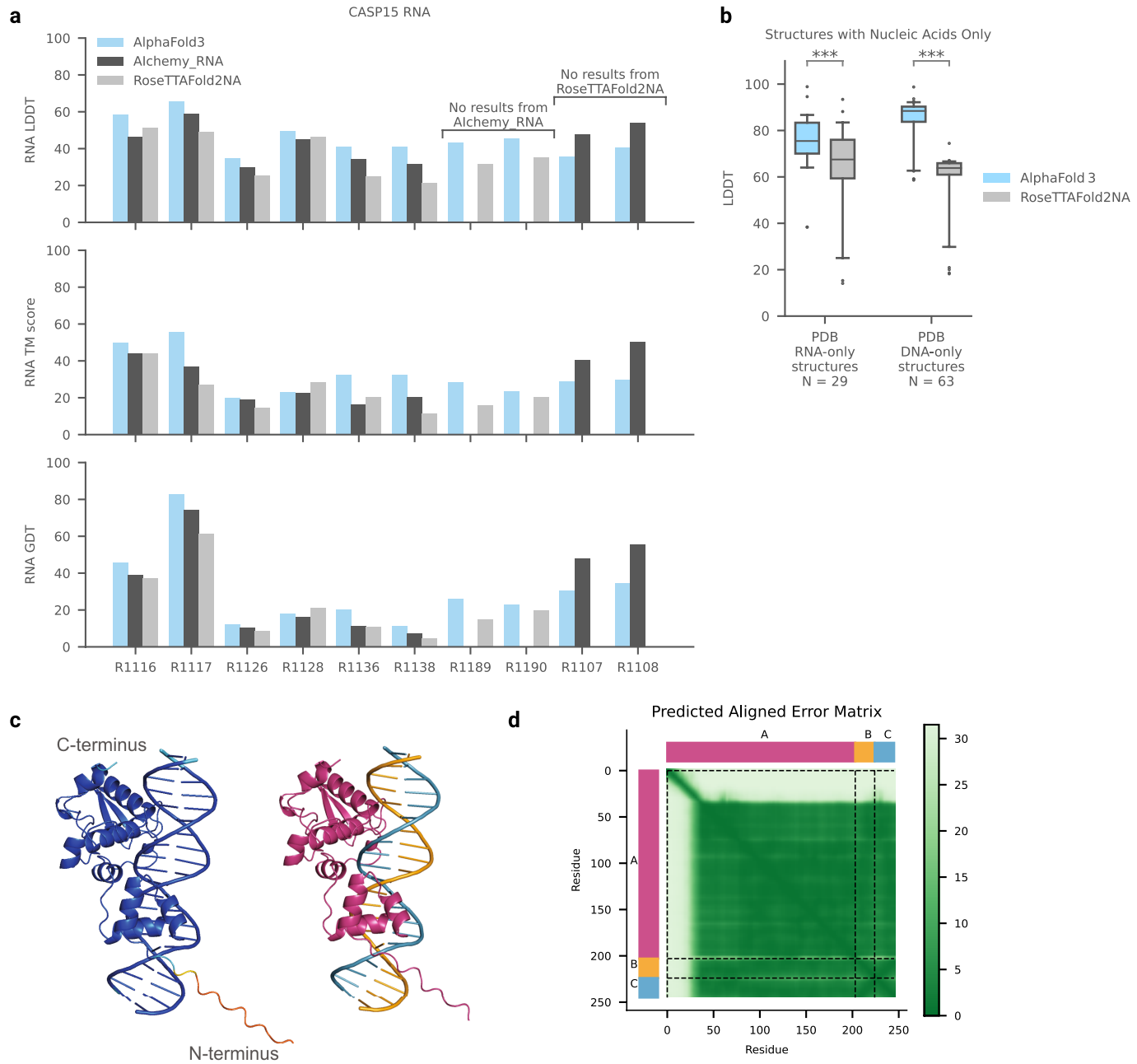
Extended Data Fig. 2 | Accuracy across training. Training curves for initial training and fine tuning showing LDDT (local distance difference test) on our evaluation set as a function of optimizer steps. One optimizer step uses a mini batch of 256 trunk samples and during initial training 256 * 48 = 12,288 diffusion samples. For fine tuning the number of diffusion samples is reduced to 256 * 32 = 8,192. The scatter plot shows the raw data points and the lines show the smoothed performance using a median filter with a kernel width of 9 data points. The dashed lines mark the points where the smoothed performance passes 90% and 97% of the initial training maximum for the first time.



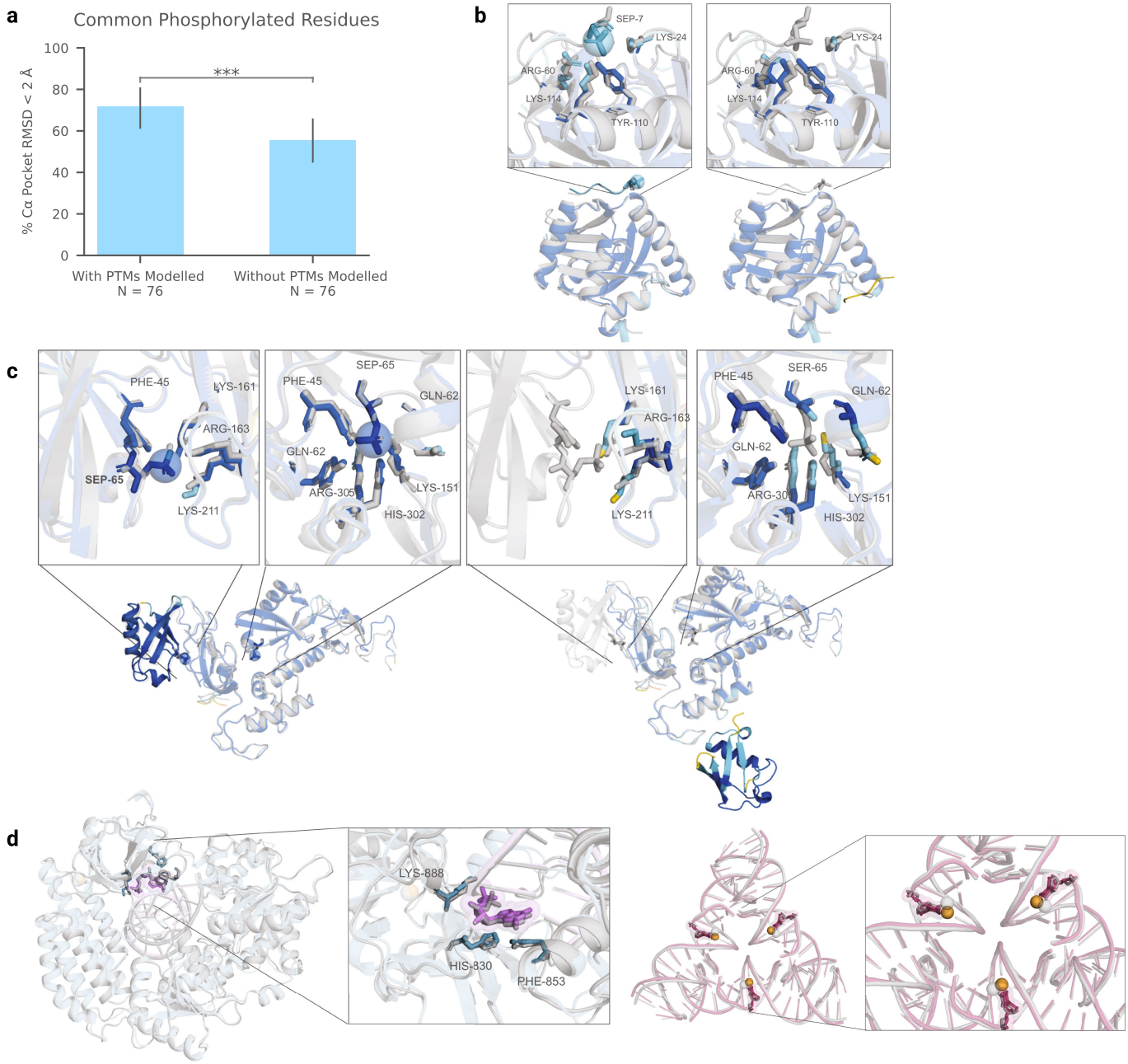
Extended Data Fig. 3 | AlphaFold 3 predictions of PoseBusters examples for which Vina and Gold were inaccurate. Predicted protein chains are shown in blue, predicted ligands in orange, and ground truth in grey. **a**, Human Notum bound to inhibitor ARUK3004556 (PDB ID 8BTI, ligand RMSD: 0.65 Å). **b**, Pseudomonas sp. PDC86 Aapf bound to HEHEAA (PDB ID 7KZ9, ligand RMSD: 1.3 Å). **c**, Human Galectin-3 carbohydrate-recognition domain in complex with compound 22 (PDB ID 7XFA, ligand RMSD: 0.44 Å).



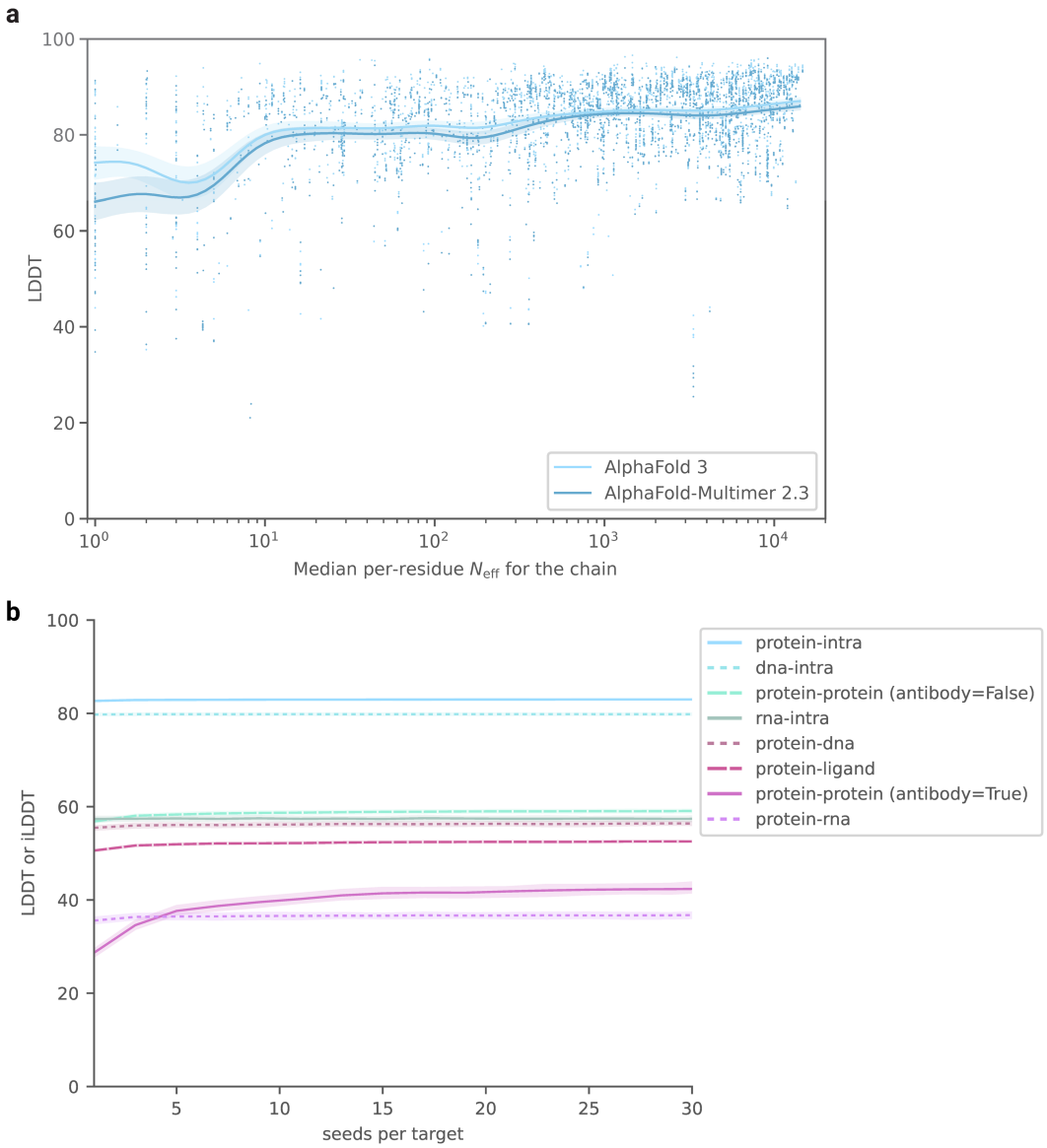
Extended Data Fig. 4 | PoseBusters analysis. a, Comparison of AlphaFold 3 and baseline method protein-ligand binding success on the PoseBusters Version 1 benchmark set (V1, August 2023 release). Methods classified by the extent of ground truth information used to make predictions. Note all methods that use pocket residue information except for UMol and AF3 also use ground truth holo protein structures. **b**, PoseBusters Version 2 (V2, November 2023 release) comparison between the leading docking method Vina and AF3 2019 (two-sided Fisher exact test, N = 308 targets, $p = 2.3 * 10^{-8}$). **c**, PoseBusters V2 results of AF3 2019 on targets with low, moderate, and high protein sequence homology (integer ranges indicate maximum sequence identity with proteins in the training set). **d**, PoseBusters V2 results of AF3 2019 with ligands split by those characterized as "common natural" ligands and others. "Common natural" ligands are defined as those which occur greater than 100 times in the PDB and which are not non-natural (by visual inspection). A full list may be found in Supplementary Table 15. Dark bar indicates RMSD < 2 Å and passing PoseBusters validity checks (PB-valid). **e**, PoseBusters V2 structural accuracy and validity. Dark bar indicates RMSD < 2 Å and passing PoseBusters validity checks (PB-valid). Light hashed bar indicates RMSD < 2 Å but not PB valid. **f**, PoseBusters V2 detailed validity check comparison. Error bars indicate exact binomial distribution 95% confidence intervals. N = 427 targets for RoseTTAFold All-Atom and 428 targets for all others in Version 1; 308 targets in Version 2.



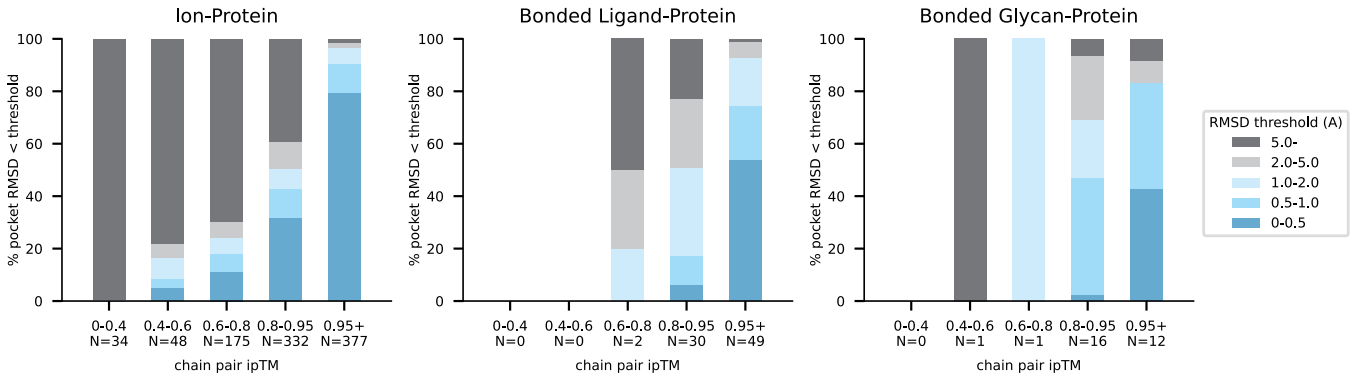
Extended Data Fig. 5 | Nucleic acid prediction accuracy and confidences. **a**, CASP15 RNA prediction accuracy from AIChemy_RNA (the top AI-based submission), RoseTTAFold2NA (the AI-based method capable of predicting proteinRNA complexes), and AlphaFold 3. Ten of the 13 targets are available in the PDB or via the CASP15 website for evaluation. Predictions are downloaded from the CASP website for external models. **b**, Accuracy on structures containing low homology RNA-only or DNA-only complexes from the recent PDB evaluation set. Comparison between AlphaFold 3 and RoseTTAFold2NA (RF2NA) (RNA: N = 29 structures, paired Wilcoxon signed-rank test, $p = 1.6 * 10^{-7}$; DNA: N = 63 structures, paired two-sided Wilcoxon signed-rank test, $p = 5.2 * 10^{-12}$). Note RF2NA was only trained and evaluated on duplexes (chains forming at least 10 hydrogen bonds), but some DNA structures in this set may not be duplexes. Box, centerline, and whiskers boundaries are at (25%, 75%) intervals, median, and (5%, 95%) intervals. **c** Predicted structure of a mycobacteriophage immunity repressor protein bound to double stranded DNA (PDB ID 7R6R), coloured by pLDDT (left; orange: 0–50, yellow: 50–70, cyan 70–90, and blue 90–100) and chain id (right). Note the disordered N-terminus not entirely shown. **d**, Predicted aligned error (PAE) per token-pair for the prediction in c with rows and columns labelled by chain id and green gradient indicating PAE.



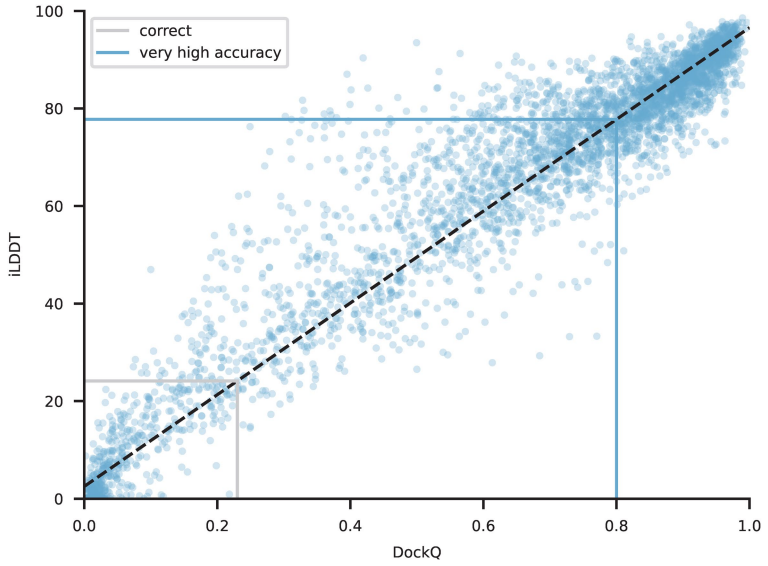
Extended Data Fig. 6 | Analysis and examples for modified proteins and nucleic acids. a, Accuracy on structures. containing common phosphorylation residues (SEP, TPO, PTR, NEP, HIP) from the recent PDB evaluation set. Comparison between AlphaFold 3 with phosphorylation modelled, and AlphaFold 3 without modelling phosphorylation (N = 76 clusters, paired two-sided Wilcoxon signed-rank test, $p = 1.6 * 10^{-4}$). Note, to predict a structure without modelling phosphorylation, we predict the parent (standard) residue in place of the modification. AlphaFold 3 generally achieves better backbone accuracy when modelling phosphorylation. Error bars indicate exact binomial distribution 95% confidence intervals. **b**, SPOC domain of human SHARP in complex with phosphorylated RNA polymerase II C-terminal domain (PDB ID 7Z1K), predictions coloured by pLDDT (orange: 0–50, yellow: 50–70, cyan 70–90, and blue 90–100). Left: Phosphorylation modelled (mean pocket-aligned $RMSD_{C\alpha}$ 2.104 Å). Right: Without modelling phosphorylation (mean pocketaligned $RMSD_{C\alpha}$ 10.261 Å). When excluding phosphorylation, AlphaFold 3 provides lower pLDDT confidence on the phosphopeptide. **c**, Structure of parkin bound to two phospho-ubiquitin molecules (PDB ID 7US1), predictions similarly coloured by pLDDT. Left: Phosphorylation modelled (mean pocket-aligned $RMSD_{C\alpha}$ 0.424 Å). Right: Without modelling phosphorylation (mean pocket-aligned $RMSD_{C\alpha}$ 9.706 Å). When excluding phosphorylation, AlphaFold 3 provides lower pLDDT confidence on the interface residues of the incorrectly predicted ubiquitin. **d**, Example structures with modified nucleic acids. Left: Guanosine monophosphate in RNA (PDB ID 7TNZ, mean pocket-aligned modified residue RMSD 0.840 Å). Right: Methylated DNA cytosines (PDB ID 7SDW, mean pocket-aligned modified residue RMSD 0.502 Å). Welabel residues of the predicted structure for reference. Ground truth structure in grey; predicted protein in blue, predicted RNA in purple, predicted DNA in magenta, predicted ions in orange, with predicted modifications highlighted via spheres.



Extended Data Fig. 7 | Model accuracy with MSA size and number of seeds. **a**, Effect of MSA depth on protein prediction accuracy. Accuracy is given as single chain LDDT score and MSA depth is computed by counting the number of non-gap residues for each position in the MSA using the N_{eff} weighting scheme and taking the median across residues (see Methods for details on N_{eff}). MSA used for AF-M 2.3 differs slightly from AF3; the data uses the AF3 MSA depth for both to make the comparison clearer. The analysis uses every protein chain in the low homology Recent PDB set, restricted to chains in complexes with fewer than 20 protein chains and fewer than 2,560 tokens (see Methods for details on Recent PDB set and comparisons to AF-M 2.3). The curves are obtained through Gaussian kernel average smoothing (window size is 0.2 units in log10(N_{eff})); the shaded area is the 95% confidence interval estimated using bootstrap of 10,000 samples. **b**, Increase in ranked accuracy with number of seeds for different molecule types. Predictions are ranked by confidence, and only the most confident per interface is scored. Evaluated on the low homology recent PDB set, filtered to less than 1,536 tokens. Number of clusters evaluated: dna-intra = 386, protein-intra = 875, rnaintra = 78, protein-dna = 307, protein-rna = 102, protein-protein (antibody = False) = 697, protein-protein (antibody = True) = 58. Confidence intervals are 95% bootstraps over 1,000 samples.



Extended Data Fig. 8 | Relationship between confidence and accuracy for protein interactions with ions, bonded ligands and bonded glycans. Accuracy is given as the percentage of interface clusters under various pocket-aligned RMSD thresholds, as a function of the chain pair ipTM of the interface. The ions group includes both metals and nonmetals. N values report the number of clusters in each band. For a similar analysis on general ligand-protein interfaces, see Fig. 4 of main text.



Extended Data Fig. 9 | Correlation of DockQ and iLDDT for protein-protein interfaces. One data point per cluster, 4,182 clusters shown. Line of best fit with a Huber regressor with epsilon 1. DockQ categories correct (>0.23), and very high accuracy (>0.8) correspond to iLDDTs of 23.6 and 77.6 respectively.

Extended Data Table 1 | Prediction accuracy across biomolecular complexes

Task	Dataset	Metric	Notes	Method	N	Mean	95% CI
Ligands	PoseBusters V1	% RMSD < 2 Å	–	RoseTTAFold All-Atom	427	42.0	37.2 – 46.8
				AF3 (2019 cutoff)	428	**76.4**	72.1 – 80.3
			Holo protein struct. given	EquiBind	428	2.6	1.3 – 4.6
				TankBind	428	15.0	11.7 – 18.7
				DiffDock	428	37.9	33.2 – 42.6
			Pocket residues specified	Vina on AF-M 2.3	428	13.1	10.0 – 16.7
				DeepDock	428	17.8	14.3 – 21.7
				Uni-Mol	428	22.9	19.0 – 27.2
				UMol	428	45.0	40.3 – 49.9
				Gold	428	51.2	46.3 – 56.0
				Vina	428	52.3	47.5 – 57.2
				Uni-Mol Docking V2	428	77.6	73.3 – 81.4
				AF3 (2019 cutoff) pocket specified	428	**90.2**	87.0 – 92.8
	PoseBusters V2	% RMSD < 2 Å	–	AF3 (2019 cutoff)	308	**80.5**	75.6 – 84.8
			Holo protein struct. given	EquiBind	308	1.9	0.7 – 4.2
				TankBind	308	15.9	12.0 – 20.5
				DiffDock	308	38.0	32.5 – 43.7
			Pocket residues specified	Vina on AF-M 2.3	308	15.3	11.4 – 19.8
				DeepDock	308	19.5	15.2 – 24.4
				Uni-Mol	308	21.8	17.3 – 26.8
				Gold	308	58.1	52.4 – 63.7
				Vina	308	59.7	54.0 – 65.3
				AF3 (2019 cutoff) pocket specified	308	**93.2**	89.8 – 95.7
Nucleic Acids	Protein-RNA	iLDDT		RoseTTAFold2NA	25	19.0	15.6 – 23.2
				AF3	25	**39.4**	28.5 – 51.9
	Protein-dsDNA	iLDDT		RoseTTAFold2NA	38	28.3	20.7 – 37.5
				AF3	38	**64.8**	56.4 – 71.7
	CASP 15 RNA	RNA LDDT		RoseTTAFold2NA	8	35.5	28.3 – 43.8
				AF3	8	47.3	41.7 – 55.2
				Alchemy_RNA2 (has human input)	8	**54.5**	45.3 – 62.4
				RNApolis (has human input)	8	50.5	45.2 – 55.8
				Chen (has human input)	8	49.8	40.7 – 58.5
				Kiharalab	8	40.9	35.1 – 54.3
				UltraFold	8	37.8	32.5 – 45.0
Covalent Mod.	Bonded ligands	% RMSD < 2 Å		AF3	66	**78.5**	68.3 – 86.2
	Glycosylation	% RMSD < 2 Å	high-quality, single-residue	AF3	28	**72.1**	53.1 – 85.7
			all-quality, single-residue	AF3	167	**46.0**	40.0 – 52.1
			all-quality, multi-residue	AF3	131	**42.4**	35.4 – 49.3
	Modified residues	% RMSD < 2 Å		AF3	154	**59.9**	52.4 – 67.0
	Modified protein residues	% RMSD < 2 Å		AF3	40	**51.0**	36.0 – 65.6
	Modified DNA residues	% RMSD < 2 Å		AF3	91	**68.6**	59.0 – 76.9
	Modified RNA residues	% RMSD < 2 Å		AF3	23	**40.9**	23.4 – 59.9
Proteins	All Protein-Protein	% dockq > 0.23		AF-M 2.3	1064	67.5	64.7 – 70.1
				AF3	1064	**76.6**	74.0 – 78.9
	Protein-Antibody	% dockq > 0.23		AF-M 2.3	65	29.6	19.6 – 40.4
				AF3	65	**62.9**	51.4 – 73.5
	Monomers	LDDT		AF-M 2.3	338	85.5	84.7 – 86.1
				AF3	338	**86.9**	86.2 – 87.6

AlphaFold 3 Performance on PoseBusters V1 (August 2023 release), PoseBusters V2 (November 6th 2023 release), and our Recent PDB evaluation set. For ligands and nucleic acids N indicates number of structures; for covalent modifications and proteins N indicates number of clusters.

Article

De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8

Received: 14 December 2022

Accepted: 7 July 2023

Published online: 11 July 2023

Open access

Check for updates

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14] ✉

There has been considerable recent progress in designing new proteins using deep-learning methods[1–9]. Despite this progress, a general deep-learning framework for protein design that enables solution of a wide range of design challenges, including de novo binder design and design of higher-order symmetric architectures, has yet to be described. Diffusion models[10,11] have had considerable success in image and language generative modelling but limited success when applied to protein modelling, probably due to the complexity of protein backbone geometry and sequence–structure relationships. Here we show that by fine-tuning the RoseTTAFold structure prediction network on protein structure denoising tasks, we obtain a generative model of protein backbones that achieves outstanding performance on unconditional and topology-constrained protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding and symmetric motif scaffolding for therapeutic and metal-binding protein design. We demonstrate the power and generality of the method, called RoseTTAFold diffusion (RFdiffusion), by experimentally characterizing the structures and functions of hundreds of designed symmetric assemblies, metal-binding proteins and protein binders. The accuracy of RFdiffusion is confirmed by the cryogenic electron microscopy structure of a designed binder in complex with influenza haemagglutinin that is nearly identical to the design model. In a manner analogous to networks that produce images from user-specified inputs, RFdiffusion enables the design of diverse functional proteins from simple molecular specifications.

De novo protein design seeks to generate proteins with specified structural and/or functional properties, for example, making a binding interaction with a given target[12], folding into a particular topology[13] or containing a catalytic site[4]. Denoising diffusion probabilistic models (DDPMs), a powerful class of machine learning models recently demonstrated to generate new photorealistic images in response to text prompts[14,15], have several properties well suited to protein design. First, DDPMs generate highly diverse outputs, as they are trained to denoise data (for instance, images or text) that have been corrupted with Gaussian noise. By learning to stochastically reverse this corruption, diverse outputs closely resembling the training data are generated. Second, DDPMs can be guided at each step of the iterative generation process towards specific design objectives through provision of conditioning information. Third, for almost all protein design applications it is necessary to explicitly model three-dimensional (3D) structures; rotationally equivariant DDPMs can do this in a global representation frame independent manner. Recent work has adapted DDPMs for protein monomer design by conditioning on small protein 'motifs'[5,9] or on secondary structure and block-adjacency ('fold') information[8]. Although promising, these attempts have shown limited success in generating sequences that fold to the intended structures in silico[5,16], probably due to the limited ability of the denoising networks to generate realistic protein backbones, and have not been tested experimentally.

We reasoned that improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure implicit in powerful structure prediction methods

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Graduate Program in Molecular Engineering, University of Washington, Seattle, WA, USA. [4]Columbia University, Department of Statistics, New York, NY, USA. [5]Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA. [6]Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [8]National Centre for Scientific Research, École Normale Supérieure rue d'Ulm, Paris, France. [9]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA, USA. [10]Department of Engineering, University of Cambridge, Cambridge, UK. [11]Faculty of Applied Sciences, Harvard University, Cambridge, MA, USA. [12]John Harvard Distinguished Science Fellowship, Harvard University, Cambridge, MA, USA. [13]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [14]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [15]These authors contributed equally: Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern. ✉e-mail: dabaker@uw.edu

such as AlphaFold2 (ref. 17) (AF2) and RoseTTAFold[18] (RF). RF has properties well suited for use in a protein design DDPM (Fig. 1a): it generates protein structures with high precision, operates on a rigid-frame representation of residues with rotational equivariance and has an architecture enabling conditioning on design specifications at the individual residue, inter-residue distance and orientation, and 3D coordinate levels. In previous work, we fine-tuned RF to complete protein backbones around input functional motifs in a single step (RF_{joint} Inpainting[4]). Experimental characterization showed that the method can scaffold a wide range of protein functional motifs with atomic accuracy[19], but the approach fails on minimalist site descriptions that do not sufficiently constrain the overall fold and, because it is deterministic, can produce only a limited diversity of designs for a given problem. We reasoned that by fine-tuning RF as the denoising network in a generative diffusion model instead, we could overcome both problems: because the starting point is random noise, each denoising trajectory yields a different solution, and because structure is built up progressively through many denoising iterations, little to no starting structural information should be required. In this study, we used an updated version of RF[18] as the basis for the denoising network architecture (Supplementary Methods), but other equivariant structure prediction networks (AF2 (ref. 17), OmegaFold[20], ESMFold[21]) could in principle be substituted into an analogous DDPM.

We construct a RF-based diffusion model, RFdiffusion, using the RF frame representation that comprises a Cα coordinate and N-Cα-C rigid orientation for each residue. We generate training inputs by noising structures sampled from the Protein Data Bank (PDB) for up to 200 steps[22]. For translations, we perturb Cα coordinates with 3D Gaussian noise. For residue orientations, we use Brownian motion on the manifold of rotation matrices (building on refs. 23,24). To enable RFdiffusion to learn to reverse each step of the noising process, we train the model by minimizing a mean-squared error (m.s.e.) loss between frame predictions and the true protein structure (without alignment), averaged across all residues (Supplementary Methods). This loss drives denoising trajectories to match the data distribution at each timestep and hence to converge on structures of designable protein backbones (Extended Data Fig. 2a). The m.s.e. contrasts to the loss used in RF structure prediction training (frame aligned point error or FAPE) in that, unlike FAPE, m.s.e. loss is not invariant to the global reference frame and therefore promotes continuity of the global coordinate frame between timesteps (Supplementary Methods).

To generate a new protein backbone, we first initialize random residue frames and RFdiffusion makes a denoised prediction. Each residue frame is updated by taking a step in the direction of this prediction with some noise added to generate the input to the next step. The nature of the noise added and the size of this reverse step is chosen such that the denoising process matches the distribution of the noising process (Supplementary Methods and Extended Data Fig. 2a). RFdiffusion initially seeks to match the full breadth of possible protein structures compatible with the purely random frames with which it is initialized, and hence the denoised structures do not initially seem protein-like (Fig. 1c, left). However, through many such steps, the breadth of possible protein structures from which the input could have arisen narrows and RFdiffusion predictions come to closely resemble protein structures (Fig. 1c, right). We use the ProteinMPNN network[1] to subsequently design sequences encoding these structures, typically sampling eight sequences per design in line with previous work[5,16] (but see Supplementary Fig. 2a). We also considered simultaneously designing structure and sequence within RFdiffusion, but given the excellent performance of combining ProteinMPNN with the diffusion of structure alone, we did not extensively explore this possibility.

Figure 1a highlights the similarities between RF structure prediction and an RFdiffusion denoising step: in both cases, the networks transform coordinates into a predicted structure, conditioned on inputs to the model. In RF, sequence is the primary input, with extra structural information provided as templates and initial coordinates to the model. In RFdiffusion, the primary input is the noised coordinates from the previous step. For specific design tasks, a range of auxiliary conditioning information, including partial sequence, fold information or fixed functional-motif coordinates can be provided (Fig. 1b and Supplementary Methods).

We explored two different strategies for training RFdiffusion: (1) in a manner akin to 'canonical' diffusion models, with predictions at each timestep independent of predictions at previous timesteps (as in previous work[5,8,9,16]), and (2) with self-conditioning[25], in which the model can condition on previous predictions between timesteps (Fig. 1a, bottom row and Supplementary Methods). The latter strategy was inspired by the success of 'recycling' in AF2, which is also central to the more recent RF model used here (Supplementary Methods). Self-conditioning within RFdiffusion notably improved performance on in silico benchmarks encompassing both conditional and unconditional protein design tasks (Fig. 2e and Extended Data Fig. 1e). Increased coherence of predictions within self-conditioned trajectories may, at least in part, explain these performance increases (Extended Data Fig. 1h). Fine-tuning RFdiffusion from pretrained RF weights was far more successful than training for an equivalent length of time from untrained weights (Extended Data Fig. 1f,g, also Supplementary Fig. 1) and the m.s.e. loss was also crucial for unconditional generation (Extended Data Fig. 1d). For all in silico benchmarks in this paper, we use the AF2 structure prediction network[17] for validation and define an in silico 'success' as an RFdiffusion output for which the AF2 structure predicted from a single sequence is (1) of high confidence (mean predicted aligned error (pAE), less than five), (2) globally within a 2 Å backbone root mean-squared deviation (r.m.s.d.) of the designed structure and (3) within 1 Å backbone r.m.s.d. on any scaffolded functional site (Supplementary Methods). This measure of in silico success has been found to correlate with experimental success[4,7,26] and is significantly more stringent than template modelling (TM)-score-based metrics used elsewhere[5,16,27–29] (Supplementary Fig. 2c,d).

Unconditional protein monomer generation

As shown in Fig. 2a–c and Supplementary Fig. 3c,d, starting from random noise, RFdiffusion can readily generate elaborate protein structures with little overall structural similarity to structures seen during training, indicating considerable generalization beyond the PDB (see Supplementary Table 1 for a comparison of all designs in the paper to the PDB). The designs are diverse (Supplementary Fig. 3a), spanning a wide range of alpha, beta and mixed alpha–beta topologies, with AF2 and ESMFold (Fig. 2c, Extended Data Fig. 1b,c and Supplementary Fig. 2b) predictions very close to the design structure models for de novo designs with as many as 600 residues. RFdiffusion generates plausible structures for even very large proteins, but these are difficult to validate in silico as they are probably generally beyond the single sequence prediction capabilities of AF2 and ESMFold. The quality and diversity of designs that are sampled are inherent to the model, and do not depend on any auxiliary conditioning input (for example, secondary structure information[8]). We experimentally characterized six of the 300 amino acid designs and three of the 200 amino acid designs, and found that they have circular dichroism spectra consistent with the mixed alpha–beta topologies of the designs and are extremely thermostable (Extended Data Fig. 3). Physics-based protein design methodologies have struggled in unconstrained generation of diverse protein monomers because of the difficulty of sampling on the very large and rugged conformational landscape[30], and overcoming this limitation has been a primary test of deep-learning based protein design approaches[5,6,8,16,27,31]. RFdiffusion strongly outperforms (based on the AF2 success metric described above) Hallucination with RF, an experimentally validated method using Monte Carlo search or gradient descent to identify sequences predicted to fold into stable structures

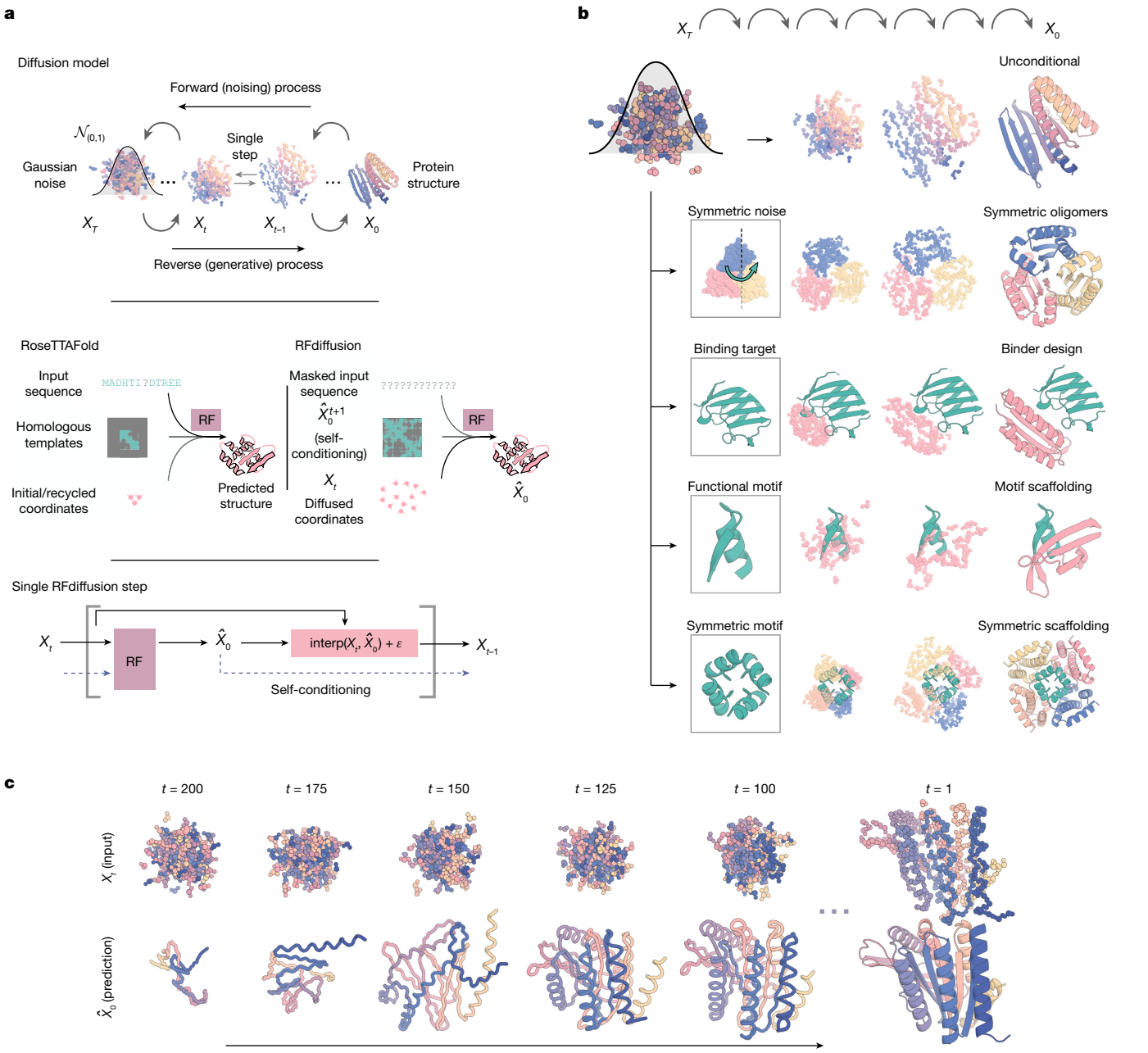


Fig. 1 | Protein design using RFdiffusion. a, Diffusion models for proteins are trained to recover corrupted (noised) protein structures and to generate new structures by reversing the corruption process through iterative denoising of initially random noise X_T into a realistic structure X_0 (top panel). The RF structure prediction network (middle panel, left side) is fine-tuned with minimal architectural changes into RFdiffusion (middle panel, right side); the denoising network of a DDPM is also shown. In RF, the primary input to the model is the sequence. In RFdiffusion, the primary input is diffused residue frames (coordinates and orientations). In both cases, the model predicts final 3D coordinates (denoted $\hat{X}_0$ in RFdiffusion). The bottom panel shows that in RFdiffusion, the model receives its previous prediction as a template input ('self-conditioning', Supplementary Methods). At each timestep t of a trajectory (typically 200 steps), RFdiffusion takes $\hat{X}_0^{t+1}$ from the previous step and X_t and then predicts an updated X_0 structure ($\hat{X}_0^t$). The next coordinate input to the model (X_{t-1}) is generated by a noisy interpolation (interp) towards $\hat{X}_0^t$. **b**, RFdiffusion is broadly applicable for protein design. RFdiffusion generates protein structures either without further input (top row) or by conditioning on (top to bottom): symmetry specifications; binding targets; protein functional motifs or symmetric functional motifs. In each case random noise, along with conditioning information, is input to RFdiffusion, which iteratively refines that noise until a final protein structure is designed. **c**, An example of an unconditional design trajectory for a 300-residue chain, depicting the input to the model (X_t) and the corresponding $\hat{X}_0$ prediction. At early timesteps (high t), $\hat{X}_0$ bears little resemblance to a protein but is gradually refined into a realistic protein structure.

(Fig. 2d). RFdiffusion generation is also more compute efficient than unconstrained Hallucination with RF, and efficiency can be greatly improved by taking larger steps at inference time and by truncating trajectories early, which is possible because RF predicts the final structure at each timestep (Extended Data Fig. 2b,c). For example, a 100-residue protein can be generated in as little as 11 s on an NVIDIA RTX A4000 Graphical Processing Unit, in contrast to RF Hallucination, which takes around 8.5 min.

It is often desirable to be able to specify a protein fold during design (such as triose-phosphate isomerase (TIM) barrels or cavity-containing

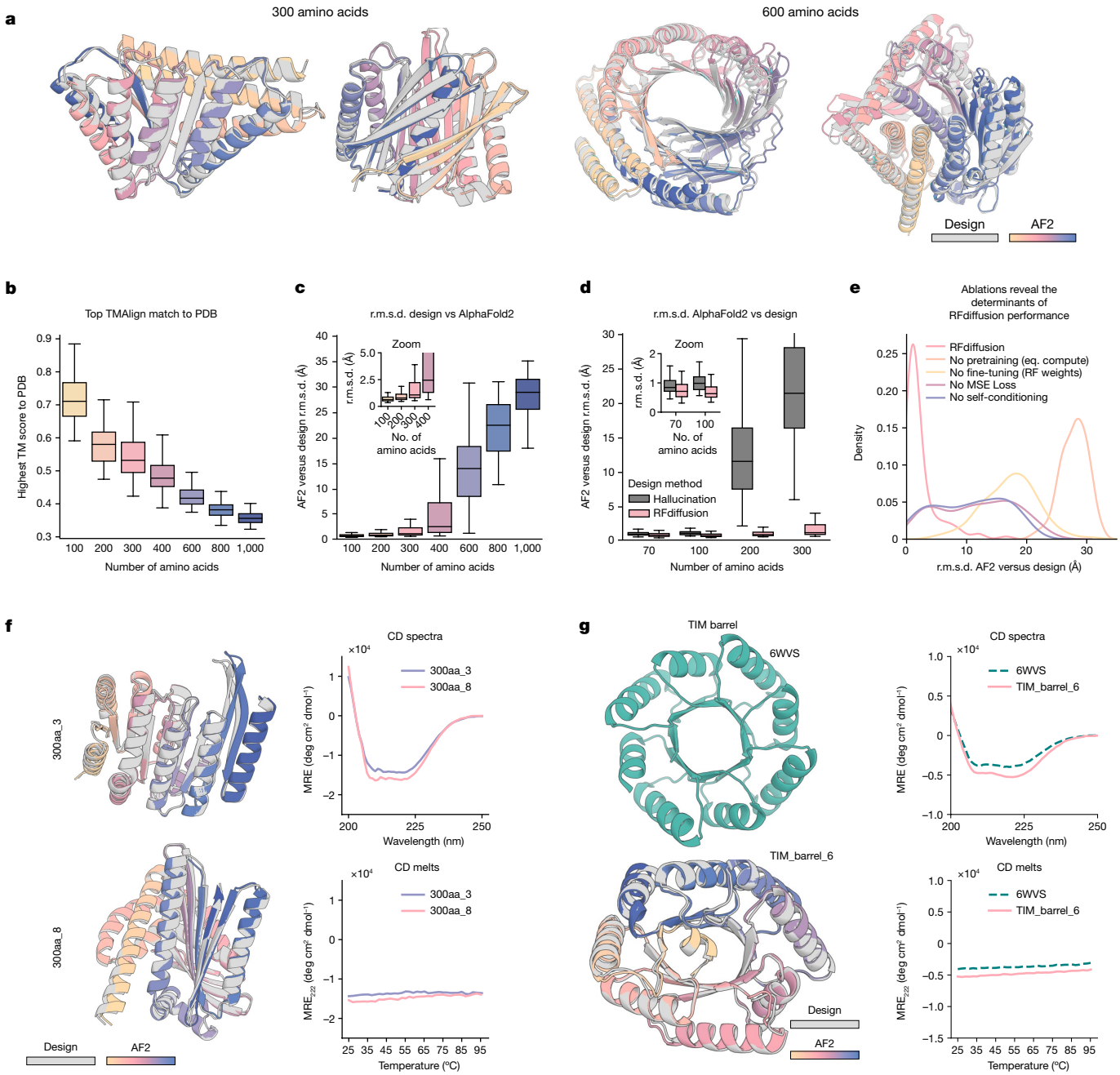


Fig. 2 | Outstanding performance of RFdiffusion for monomer generation. **a**, RFdiffusion can generate new monomeric proteins of different lengths (left 300, right 600) with no conditioning information. Grey, design model; colours, AF2 prediction. r.m.s.d. AF2 versus design (Å), left to right: 0.90, 0.98, 1.15, 1.67. **b**, Unconditional designs from RFdiffusion are new and not present in the training set as quantified by highest TM-score to the PDB; the divergence from previously known structures increases with length. **c**, Unconditional samples are closely repredicted by AF2 up to about 400 amino acids. **d**, RFdiffusion significantly outperforms Hallucination (with RF) at unconditional monomer generation (two-proportion z-test of in silico success: $n = 400$ designs per condition, $z = 9.5$, $P = 1.6 \times 10^{-21}$). Although Hallucination successfully generates designs up to 100 amino acids in length, in silico success rates rapidly deteriorate beyond this length. **e**, Ablating pretraining (by starting from untrained RF), RFdiffusion fine-tuning (that is, using original RF structure prediction weights as the denoiser), self-conditioning or m.s.e. losses (by training with FAPE) each notably decrease the performance of RFdiffusion. r.m.s.d. between design and AF2 is shown, for the unconditional generation of 300 amino acid proteins (Supplementary Methods). **f**, Two example 300 amino acid proteins that expressed as soluble monomers. Designs (grey) overlaid with AF2 predictions (colours) are shown on the left, alongside circular dichroism (CD) spectra (top) and melt curves (bottom) on the right. The designs are highly thermostable. **g**, RFdiffusion can condition on fold information. An example TIM barrel is shown (bottom left), conditioned on the secondary structure and block adjacency of a previously designed TIM barrel, PDB 6WVS (top left). Designs have very similar circular dichroism spectra to PDB 6WVS (top right) and are highly thermostable (bottom right). See also Extended Data Fig. 3 for further traces. Boxplots represent median ± interquartile range; tails are minimum and maximum excluding outliers (±1.5× interquartile range).

NTF2s for small molecule binder and enzyme design[32,33]), and thus we further fine-tuned RFdiffusion to condition on secondary structure and/or fold information, enabling rapid and accurate generation of diverse designs with the desired topologies (Fig. 2g and Extended Data Fig. 4). In silico success rates were 42.5 and 54.1% for TIM barrels and NTF2 folds, respectively (Extended Data Fig. 4d), and experimental

characterization of 11 TIM barrel designs indicated that at least eight designs were soluble, thermostable and had circular dichroism spectra consistent with the design model (Fig. 2g and Extended Data Fig. 4e,f).

Design of higher-order oligomers

There is considerable interest in designing symmetric oligomers, which can serve as vaccine platforms[34], delivery vehicles[35] and catalysts[36]. Cyclic oligomers have been designed using structure prediction networks with an adaptation of Hallucination that searches for sequences predicted to fold to the desired cyclic symmetry, but this approach fails for higher-order dihedral, tetrahedral, octahedral and icosahedral symmetries, probably in part because of the much lower representation of such structures in the PDB[7].

We set out to generalize RFdiffusion to create symmetric oligomeric structures with any specified point group symmetry. Given a specification of a point group symmetry for an oligomer with n chains, and the monomer chain length, we generate random starting residue frames for a single monomer subunit as in the unconditional generation case, and then generate $n - 1$ copies of this starting point arranged with the specified point group symmetry. Because RFdiffusion is equivariant (inherited from RF) with respect to rotation and relabelings of chains, symmetry is largely maintained in the denoising predictions; we explicitly resymmetrize at each step but this changes the structures only slightly (compare grey and coloured chains in Extended Data Fig. 5a and Supplementary Methods). For octahedral and icosahedral architectures, we explicitly model only the smallest subset of monomers required to generate the full assembly (for example, for icosahedra, the subunits at the five-, three- and twofold symmetry axes) to reduce the computational cost and memory footprint.

Despite not being trained on symmetric inputs, RFdiffusion is able to generate symmetric oligomers with high in silico success rates (Extended Data Fig. 5b), particularly when guided by an auxiliary inter- and intrachain contact potential (Extended Data Fig. 5c). As illustrated in Fig. 3 and Extended Data Fig. 5e, RFdiffusion designs are nearly indistinguishable from AF2 predictions of the structures adopted by the designed sequences, and many show little resemblance to previously solved protein structures (Extended Data Fig. 5d and Supplementary Table 1). Several of the oligomeric topologies are not seen in the PDB, including two-layer beta barrels (Fig. 3a, C10 symmetry) and complex mixed alpha/beta topologies (Fig. 3a, C8 symmetry; closest TM align in PDB 6BRP, 0.47, and PDB 6BRO, 0.43, respectively).

We selected 608 designs for experimental characterization and found using size-exclusion chromatography (SEC) that at least 87 had oligomerization states closely consistent with the design models (within the 95% confidence interval, 126 designs within the 99% confidence interval, as determined by SEC calibration curves; Supplementary Figs. 4 and 5). We took advantage of the increased size of these oligomers (compared to the smaller unconditional and fold-conditioned monomers described above) and collected negative stain electron microscopy (nsEM) data on a subset of these designs across different symmetry groups. For most, distinct particles were evident with shapes resembling the design models in both the raw micrographs and subsequent two-dimensional (2D) classifications (Fig. 3 and Extended Data Fig. 5f). nsEM characterization of a C3 design (HE0822) with 350 residue subunits (1,050 residues in total) suggests that the actual structure is very close to the design, both over the 350 residue subunits and the overall C3 architecture. 2D class averages are clearly consistent with both top and side views of the design model, and a 3D reconstruction of the density has key features consistent with the design, including the distinctive pinwheel shape (Fig. 3b, top row). Electron microscopy 2D class averages of C5 and C6 designs with more than 750 residues (HE0794, HE0789, HE0841) were also consistent with the respective design models (Extended Data Fig. 5f).

RFdiffusion also generated cyclic oligomers with alpha and/or beta barrel structures that resemble expanded TIM barrels and provide an interesting comparison between innovation during natural evolution and innovation through deep learning. The TIM barrel fold, with eight strands and eight helices, is one of the most abundant folds in nature[37]. nsEM confirmed the structure of two RFdiffusion designed cyclic oligomers, which considerably extend beyond this fold (Fig. 3b, bottom rows). HE0626 is a C6 alpha–beta barrel composed of 18 strands and 18 helices, and HE0675 is a C8 octamer composed of an inner ring of 16 strands and an outer ring of 16 helices arranged locally in a very similar repeating pattern to the TIM barrel (1:1 helix:strand). For both HE0626 and HE0675 we obtained nsEM 3D reconstructions that are in agreement with the computational design models. The HE0600 design is also an alpha–beta barrel (Extended Data Fig. 5f), but has two strands for every helix (24 strands and 12 helices in total) and hence is locally different from a TIM barrel. Whereas natural evolution has extensively explored structural variations of the classic eight-strand or eight-helix TIM barrel fold, RFdiffusion can more readily explore global changes in barrel curvature, enabling discovery of TIM barrel-like structures with many more helices and strands.

RFdiffusion also readily generated structures with dihedral, tetrahedral and icosohedral symmetries (Fig. 3c,d and Extended Data Fig. 5e,f). SEC characterization indicated that 38 D2, seven D3 and three D4 designs had the expected molecular weights (these have four, six and eight chains, respectively) (Supplementary Fig. 5). Although the D2 dihedrals are too small for nsEM, 2D class averages—and for some, 3D reconstructions of D3 and D4 designs—were congruent with the overall topologies of the design models (Fig. 3c and Extended Data Fig. 5f). Similarly, 3D reconstruction (Fig. 3c) and cryogenic electron microscopy (cryo-EM) 2D class averages (Extended Data Fig. 5g and Supplementary Fig. 6) of the D4 HE0537 closely match the design model, recapitulating the roughly 45° offset between tetramic subunits. 2D nsEM class averages for a 12-chain tetrahedron (HE0964) were consistent with the design model (Extended Data Fig. 5f). Forty-eight icosahedra were selected for experimental validation, and one, HE0902, a 15 nm (diameter) highly porous assembly (Fig. 3d, left) was observed in nsEM micrographs to form homogeneous particles. 2D class averages and a 3D reconstruction very closely match the design model (Fig. 3d), with triangular hubs arrayed around the empty C5 axes. Designs such as HE0902 (and future similar large assemblies) should be useful as new nanomaterials and vaccine scaffolds, with robust assembly and (in the case of HE0902) the outward facing N and C termini offering many possibilities for antigen display.

Functional-motif scaffolding

We next investigated the use of RFdiffusion for scaffolding protein structural motifs that carry out binding and catalytic functions, in which the role of the scaffold is to hold the motif in precisely the 3D geometry needed for optimal function. In RFdiffusion, we input motifs as 3D coordinates (including sequence and sidechains) both during conditional training and inference, and build scaffolds that hold the motif atomic coordinates in place. Many deep-learning methods have been developed recently to address this problem, including RF_{joint} Inpainting[4], constrained Hallucination[4] and other DDPMs[5,8,29]. To rigorously evaluate the performance of these methods in comparison to RFdiffusion across a broad set of design challenges, we established an in silico benchmark test (Supplementary Table 9) comprising 25 motif-scaffolding design problems addressed in six recent publications encompassing several design methodologies[4,5,29,38–40]. The challenges span a broad range of motifs, including simple 'inpainting' problems, viral epitopes, receptor traps, small molecule binding sites, binding interfaces and enzyme active sites.

RFdiffusion solves 23 of the 25 benchmark problems, compared to 15 for Hallucination and 19 for RF_{joint} Inpainting (Fig. 4a,b). For 19 out

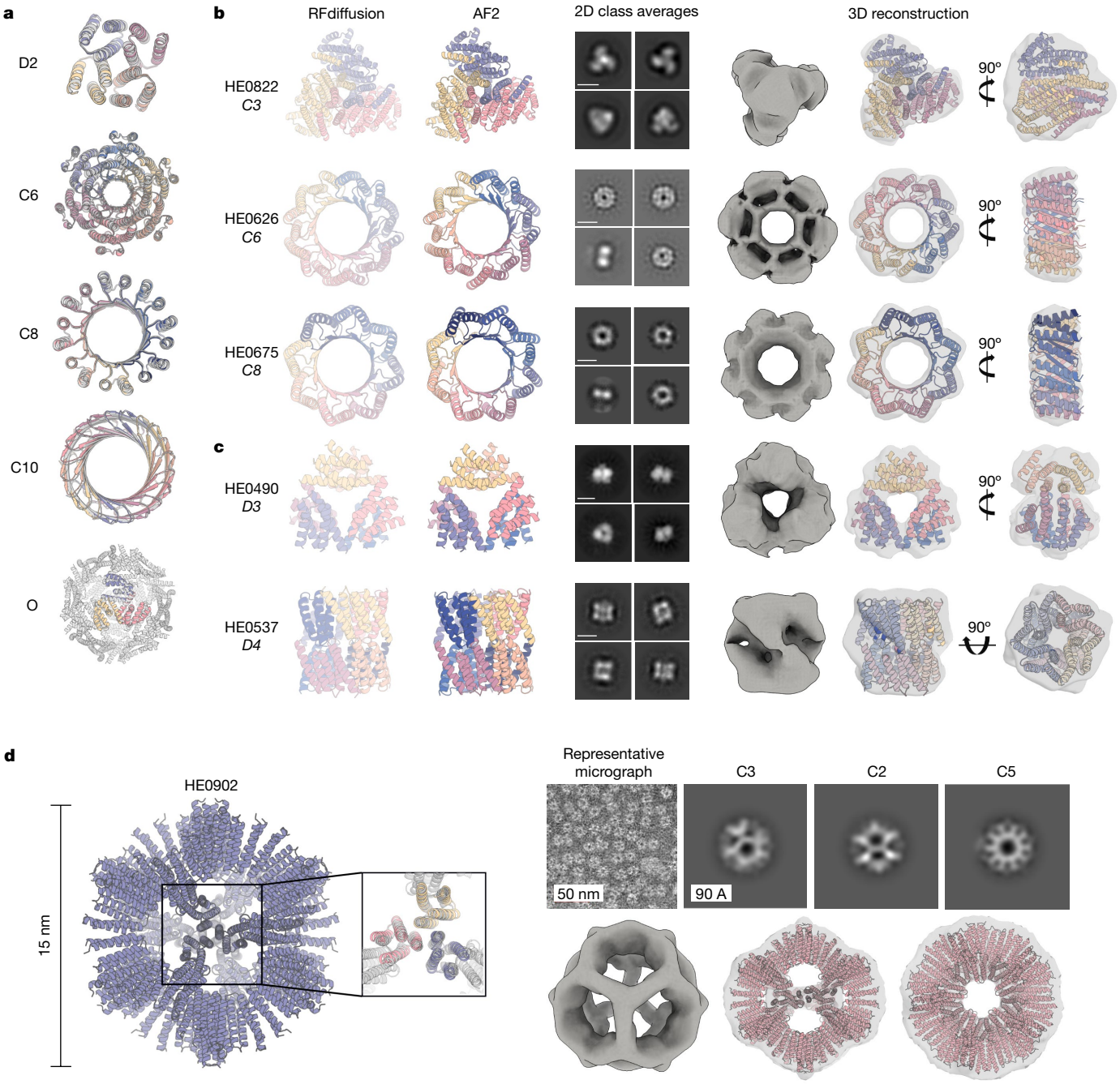


Fig. 3 | Design and experimental characterization of symmetric oligomers. **a**, RFdiffusion-generated assemblies overlaid with the AF2 structure predictions based on the designed sequences; in all five cases they are nearly indistinguishable (for the octahedron (bottom), the prediction was for the C3 substructure). Symmetries are indicated to the left of the design models. **b**,**c**, Designed assemblies characterized by nsEM. Model symmetries are as follows: cyclic, C3 (HE0822, 350 amino acids (AA) per chain), C6 (HE0626, 100 AA per chain) and C8 (HE0675, 60 AA per chain) (**b**); dihedral, D3 (HE0490, 80 AA per chain) and D4 (HE0537, 100 AA per chain) (**c**). From left to right: (1) symmetric design model, (2) AF2 prediction of design following sequence design with ProteinMPNN, (3) 2D class averages showing both top and side views (scale bar, 60 Å for all class averages) and (4) 3D reconstructions from class averages with the design model fit into the density map. The overall shapes are consistent with the design models, and confirm the intended oligomeric state. As in **a**, AF2 predictions of each design are nearly indistinguishable from the design model (backbone r.m.s.d.s (Å) for HE0822, HE0626, HE0490, HE0675 and HE0537, are 1.33, 1.03, 0.60, 0.74 and 0.75, respectively). **d**, nsEM characterization of an icosahedral particle (HE0902, 100 AA per chain). The design model, including the AF2 prediction of the C3 subunit are shown on the left. nsEM data are shown on the right: on top, a representative micrograph is shown alongside 2D class averages along each symmetry axis (C3, C2 and C5, from left to right) with the corresponding 3D reconstruction map views shown directly below overlaid on the design model.

of 23 of the problems solved by RFdiffusion, the fraction of successful designs is higher than either Hallucination or RF_{joint} Inpainting. The excellent performance of RFdiffusion required no hyperparameter tuning or external potentials; this contrasts with Hallucination, for which problem-specific optimization can be required. In 17 out of 23 of the problems, RFdiffusion-generated successful solutions with higher in silico success rates when noise was not added during the reverse diffusion trajectories (see Extended Data Fig. 1i for further discussion on the

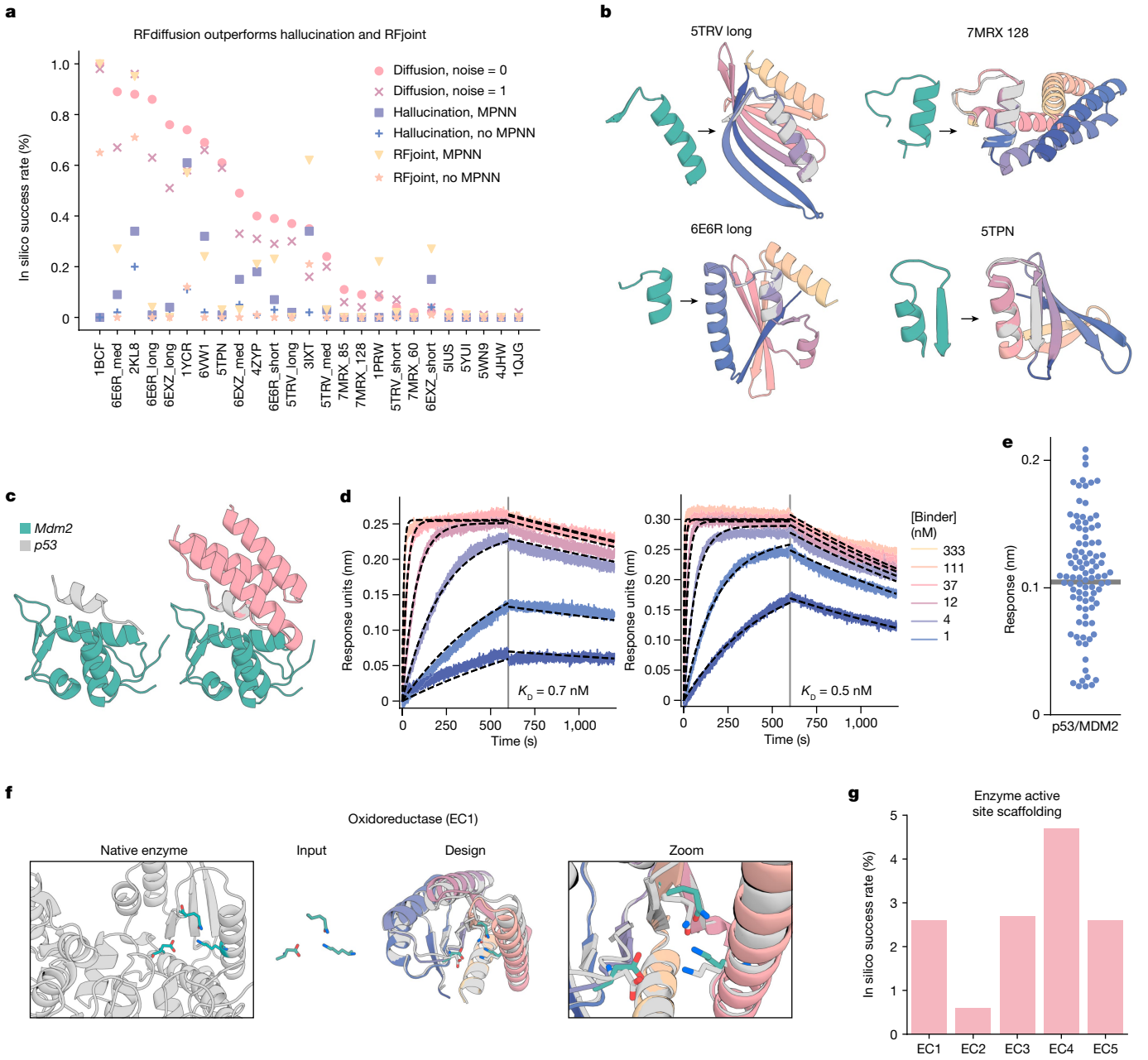


Fig. 4 | Scaffolding of diverse functional sites with RFdiffusion. a, RFdiffusion outperforms other methods across 25 benchmark motif-scaffolding problems collected from six recent publications (Supplementary Table 9). In silico success is defined as AF2 r.m.s.d. to design model less than 2 Å, AF2 r.m.s.d. to the native functional motif less than 1 Å and AF2 pAE less than five. One hundred designs were generated per problem, with no previous optimization on the benchmark set (some optimization was necessary for Hallucination). Supplementary Table 10 presents full results. In silico success rates on the problems are correlated between the methods, and RFdiffusion can still struggle on challenging problems in which all methods have low success. **b**, Four examples of designs in which RFdiffusion significantly outperforms existing methods. Teal, native motif; colours, AF2 prediction of a design. Metrics (r.m.s.d. AF2 versus design/versus native motif (Å), AF2 pAE): 5TRV long, 1.17/0.57; 4.73; 6E6R long, 0.89/0.27, 4.56; 7MRX long, 0.84/0.82 4.32; 5TPN, 0.59/0.49 3.77. **c**, RFdiffusion can scaffold the p53 helix that binds MDM2 (left) and makes extra contacts with the target (right, average 31% increased surface area. Design was p53_design_89). Designs were generated with an RFdiffusion model fine-tuned on complexes. **d**, BLI measurements indicate high-affinity binding to MDM2 (p53_design_89, 0.7 nM; p53_design_53, 0.5 nM); the native affinity is 600 nM (ref. 42). **e**, Out of 95 designs, 55 showed binding to MDM2 (more than 50% of maximum response). Thirty-two of these were monomeric (Supplementary Fig. 10h). **f**, After fine-tuning (Supplementary Methods), RFdiffusion can scaffold enzyme active sites. An oxidoreductase example (EC1) is shown (PDB 1A4I); catalytic site (teal); RFdiffusion output (grey, model; colours, AF2 prediction); zoom of active site. AF2 versus design backbone r.m.s.d. 0.88 Å, AF2 versus design motif backbone r.m.s.d. 0.53 Å, AF2 versus design motif full-atom r.m.s.d. 1.05 Å, AF2 pAE 4.47. **g**, In silico success rates on active sites derived from EC1-5 (AF2 Motif r.m.s.d. versus native: backbone less than 1 Å, backbone and sidechain atoms less than 1.5 Å, r.m.s.d. AF2 versus design less than 2 Å, AF2 pAE less than 5).

effect of noise on design quality, and Supplementary Fig. 8 for analysis of design diversity). The ability of RFdiffusion to scaffold functional motifs is not related to their presence in the RFdiffusion training set (Supplementary Fig. 7).

One of the benchmark problems is the scaffolding of the p53 helix that binds MDM2. Inhibiting this interaction through high-affinity competitive inhibition by scaffolding the p53 helix and making further interactions with MDM2 is a promising therapeutic avenue[41]. In silico

success has been described elsewhere[4], but experimental success has not been reported. We used an RFdiffusion model fine-tuned on protein complexes (Supplementary Methods) to generate 96 designs scaffolding this helix. We scaffolded the p53 helix in the presence of MDM2, so extra interactions could be designed by RFdiffusion and experimentally identified 0.5 and 0.7 nM binders (Fig. 4c,d), three orders of magnitude higher affinity than the reported 600 nM affinity of the p53 peptide alone[42]. The overall success rate was quite high: out of the 96 designs, 55 showed some detectable binding at 10 μM (Fig. 4e and Supplementary Fig. 10h).

Scaffolding enzyme active sites

A grand challenge in protein design is to scaffold minimal descriptions of enzyme active sites comprising a few single amino acids. Whereas some in silico success has been reported previously[4], a general solution that can readily produce high-quality, orthogonally validated outputs remains elusive. Following fine-tuning on a task mimicking this problem (Supplementary Methods), RFdiffusion was able to scaffold enzyme active sites comprising many sidechain and backbone functional groups with high accuracy and in silico success rates across a range of enzyme classes (Fig. 4f and Extended Data Fig. 6a–d; in silico success required fine tuning). Although RFdiffusion is unable to explicitly model bound small molecules at present (however, see our conclusions), the substrate can be implicitly modelled using an external potential to guide the generation of 'pockets' around the active site. As a demonstration, we scaffold a retroaldolase active site triad while implicitly modelling the reaction substrate (Extended Data Fig. 6e–h).

Symmetric functional-motif scaffolding

Several important design challenges involve the scaffolding of several copies of a functional motif in symmetric arrangements. For example, many viral glycoproteins are trimeric and symmetry matched arrangements of inhibitory domains can be extremely potent[43–46]. Conversely, symmetric presentation of viral epitopes in an arrangement that mimics the virus could induce new classes of neutralizing antibodies[47,48]. To explore this general direction, we sought to design trimeric multivalent binders to the SARS-CoV-2 spike protein. In previous work, flexible linkage of a binder to the ACE2 binding site (on the spike protein receptor binding domain) to a trimerization domain yielded a high-affinity inhibitor that had potent and broadly neutralizing antiviral activity in animal models[43]. Ideally, however, symmetric fusions to binders would be rigid, so as to reduce the entropic cost of binding while maintaining the avidity benefits from multivalency. We used RFdiffusion to design C3-symmetric trimers that rigidly hold three binding domains (the functional motif in this case) such that they exactly match the ACE2 binding sites on the SARS-CoV-2 spike protein trimer. The designs were confidently predicted by AF2 to both assemble as C3-symmetric oligomers, and to scaffold the AHB2 SARS-CoV-2 binder interface with high accuracy (Fig. 5a).

The ability to scaffold functional sites with any desired symmetry opens up new approaches to designing metal-coordinating protein assemblies[49,50]. Divalent transition metal ions show distinct preferences for specific coordination geometries (for example, square planar, tetrahedral and octahedral) with ion-specific optimal sidechain–metal bond lengths. RFdiffusion provides a general route to building up symmetric protein assemblies around such sites, with the symmetry of the assembly matching the symmetry of the coordination geometry. As a first test, we sought to design square-planar Ni^{2+} binding sites. We designed C4 protein assemblies with four central histidine imidazoles arranged in an ideal Ni^{2+}-binding site with square-planar coordination geometry (Fig. 5b). Diverse designs starting from distinct C4-symmetric histidine square-planar sites had good in silico success with the histidine residues in near ideal geometries for coordinating metal in the AF2-predicted structures (Supplementary Fig. 9).

We expressed and purified 44 designs in *Escherichia coli*, and found that 37 had SEC chromatograms consistent with the intended oligomeric state (Extended Data Fig. 7b). Of the designs, 36 were tested for Ni^{2+} coordination by isothermal titration calorimetry, and 18 were found to bind Ni^{2+} with dissociation constants ranging from low nanomolar to low micromolar (Fig. 5c,d and Extended Data Fig. 7a). The inflection points in the wild-type isotherms indicate binding with the designed stoichiometry, a one to four ratio of ion to monomer. Although most of the designed proteins showed exothermic metal coordination, in a few cases binding was endothermic (Fig. 5d, left and Extended Data Fig. 7a: NiB2.9, NiB2.10, NiB2.15 and NiB2.23), suggesting that Ni^{2+} coordination is entropically driven in these assemblies. To confirm that Ni^{2+} binding was indeed mediated by the scaffolded histidine 52, we mutated this residue to alanine, which abolished or notably reduced binding in 17 out of 17 cases with successful expression (Extended Data Figs. 7a,c and Fig. 5c,d; one mutant did not express). We structurally characterized by nsEM a subset of the designs—NiB1.12, NiB1.15, NiB1.17 and NiB1.20—that showed histidine-dependent binding. All four designs showed clear fourfold symmetry both in the raw micrographs and in 2D class averages (Fig. 5c,d), with design NiB1.17 also clearly showing twofold axis side views with a measured diameter approximating the design model. A 3D reconstruction of NiB1.17 was in close agreement with the design model (Fig. 5c).

Design of protein-binding proteins

The design of high-affinity binders to target proteins is a grand challenge in protein design, with numerous therapeutic applications[51]. A general method for de novo binder design from target structure information alone using the physically based Rosetta method was recently described[12], and subsequently, using ProteinMPNN for sequence design and AF2 for design filtering was found to improve design success rates[26]. However, experimental success rates were low, still requiring many thousands of designs to be screened for each design campaign[12], and the approach relied on prespecifying a particular set of protein scaffolds as the basis for the designs, inherently limiting the diversity and shape complementarity of possible solutions[12]. To our knowledge, no deep-learning method has yet demonstrated experimental general success in designing completely de novo binders.

We reasoned that RFdiffusion might be able to address this challenge by directly generating binding proteins in the context of the target. For many therapeutic applications, for example, blocking a protein–protein interaction, it is desirable to bind to a particular site on a target protein. To enable this, we fine-tuned RFdiffusion on protein complex structures, providing a feature as input indicating a subset of the residues on the target chain (called 'interface hotspots') to which the diffused chain binds (Fig. 6a and Extended Data Fig. 8a,b). For design challenges in which a particular binder fold might be especially compatible, we enabled coarse-grained control over binder scaffold topology by fine-tuning an extra model to condition binder diffusion on secondary structure and block-adjacency information, in addition to conditioning on interface hotspots (Extended Data Fig. 8c,d and Supplementary Methods).

To compare RFdiffusion to previous binder design methods, we performed binder design campaigns against five targets: Influenza A H1 Haemagglutinin (HA)[52], Interleukin-7 Receptor-α (IL-7Rα)[12], Programmed Death-Ligand 1 (PD-L1)[12], Insulin Receptor (InsR) and Tropomyosin Receptor Kinase A (TrkA)[12]. We designed putative binders to each target, both with and without conditioning on compatible fold information, with high in silico success rates (Extended Data Fig. 8e,f). Designs were filtered by AF2 confidence in the interface and monomer structure[26], and 95 were selected for each target for experimental characterization.

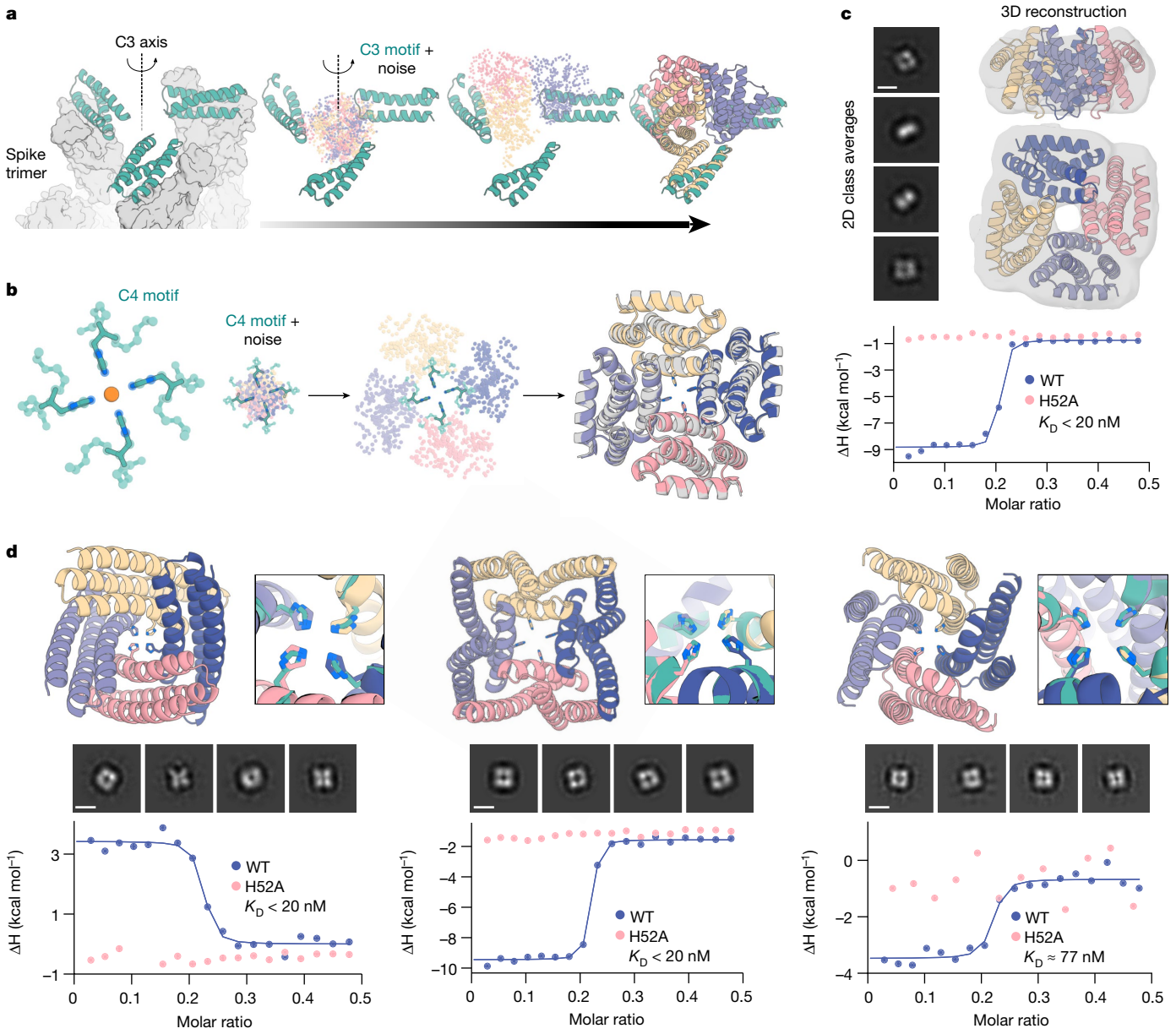


Fig. 5 | Symmetric motif scaffolding with RFdiffusion. a, Design of symmetric oligomers scaffolding the binding interface of ACE2 mimic AHB2 (left, teal) against the SARS-CoV-2 spike trimer (left, grey). Three AHB2 copies are input to RFdiffusion along with C3 noise (middle); output are C3-symmetric oligomers holding the three AHB2 copies in place to engage all spike subunits. AF2 predictions (right) recapitulate the AHB2 structure with 0.6 Å r.m.s.d. over the assymetric unit and 2.9 Å r.m.s.d. over the C3 assembly. **b**, Design of C4-symmetric oligomers to scaffold a Ni^{2+} binding motif (left). Starting from square-planar histidine rotamers within helical fragments (Supplementary Methods), RFdiffusion generates a C4 oligomer scaffolding the binding domain (middle). AF2 predictions (colour) agree closely with the design model (grey), with backbone r.m.s.d. less than 1.0 Å (right). **c**, nsEM 2D class averages (scale bar, 60 Å) and 3D reconstruction density are consistent with the symmetry and structure of the NiB1.17 design model shown superimposed on the density in ribbon representation (top). Isothermal titration calorimetry binding isotherm of design NiB1.17 (blue) indicates a dissociation constant less than 20 nM at a metal:monomer stoichiometry of 1:4. The H52A mutant isotherm (pink) ablates binding, indicating scaffolded histidine residues are critical for metal binding. **d**, Additional experimentally characterized Ni^{2+} binders NiB2.15 (left), NiB1.12 (middle) and NiB1.20 (right). Metal-coordinating sidechains in the design models (top, teal) are closely recapitulated in the AF2 predictions (colours). 2D nsEM class averages (middle; scale bar, 60 Å) are consistent with design models. Binding isotherms for wild-type (WT) and H52A mutant (bottom) indicate Ni^{2+} binding mediated directly by the scaffolded histidines at the designed stoichiometry. Note that for ITC plots, points represent single measurements.

The designed binders were expressed in *E. coli* and purified, and binding was assessed through single point biolayer interferometry (BLI) screening at 10 μM binder concentration (Extended Data Fig. 8g). The overall experimental success rate, defined as binding at or above 50% of the maximal response for the positive control, was 19% (this is a conservative estimate as some designs that showed binding had insufficient material to permit screening at 10 μM: Extended Data Fig. 8g); an increase of roughly two orders of magnitude over our previous Rosetta-based method on the same targets (Fig. 6b). Binders were identified for all five targets, with fewer than 100 designs tested per target compared to thousands in previous studies. Full BLI titrations for a subset of the designs showed nanomolar affinities with no further experimental optimization, including HA and IL-7Rα binders with affinities of roughly 30 nM (Fig. 6c). Binding

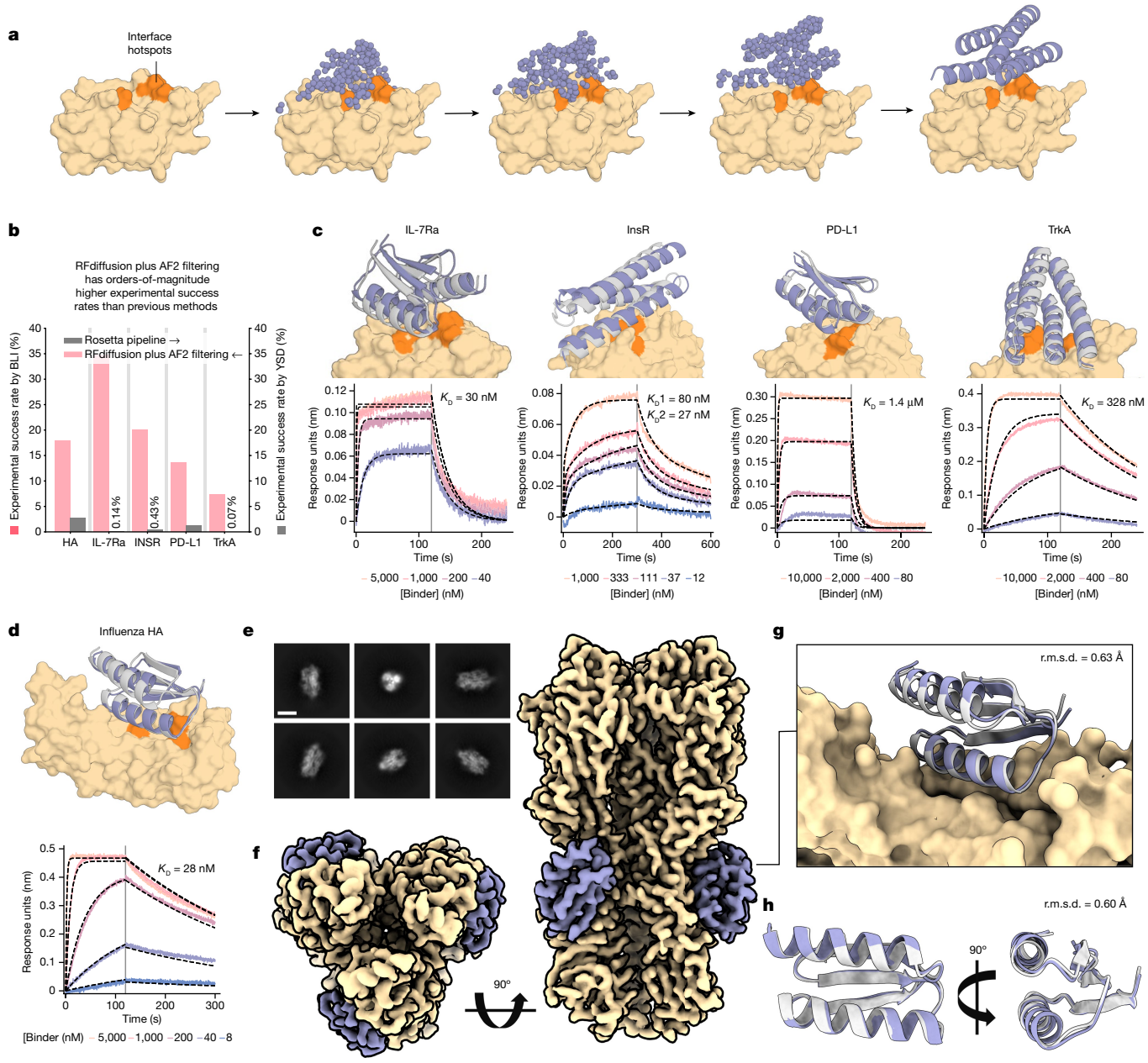


Fig. 6 | De novo design of protein-binding proteins. a, RFdiffusion generates protein binders given a target and specification of interface hotspot residues. **b**, De novo binders were designed to five protein targets; Influenza A H1 HA, IL-7Rα, InsR, PD-L1 and TrkA and hits with BLI response greater than or equal to 50% of the positive control were identified for all targets. For IL-7Rα, InsR, PD-L1 and TrkA, RFdiffusion has success rates roughly two orders of magnitude higher than the original design campaigns. We attribute one order of magnitude to RFdiffusion, and the second to filtering with AF2 (estimated success rates for previous campaigns if AF2 filtering had been used: HA, 0%; IL-7Rα, 2.2%; InsR, 5.5%; PD-L1, 3.7%; TrkA, 1.5%). **c**, For IL-7Rα, InsR, PD-L1 and TrkA, the highest affinity binder is shown above a BLI titration series. Reported K_D values are based on global kinetic fitting with fixed global R_{max}. **d**, The highest affinity HA binder, HA_20, binds with a K_D of 28 nM. **c**,**d**, Yellow or orange, target or hotspot residues; grey, design model; purple, AF2 prediction (r.m.s.d. AF2 versus design). Binders: IL7Ra_55 (2.1 Å), InsulinR_30 (2.6 Å), PDL1_77 (1.5 Å), TrkA_88 (1.4 Å) (left to right in **c**) and HA_20 (1.7 Å) (**d**). **e**, Cryo-EM 2D class averages of *HA_20* bound to influenza HA, strain A/USA:Iowa/1943 H1N1 (scale bar, 10 nm). **f**, 2.9 Å cryo-EM 3D reconstruction of the complex viewed along two orthogonal axes. *HA_20* (purple) is bound to H1 along the stem of all three subunits. **g**, The cryo-EM structure of the HA_20 binder in complex closely matches the design model (r.m.s.d. to RFdiffusion design, 0.63 Å; yellow, influenza HA). **h**, Structure of the HA_20 binder alone superimposed on the design model viewed along two orthogonal axes. For cryo-EM panels, yellow, Influenza H1 map and/or structure; grey, HA_20 binder design model; purple, HA_20 binder map or structure.

interfaces were often highly distinct from interfaces to these targets in the PDB (Supplementary Figs. 11 and 12). To assess binder specificity, six of the highest affinity IL-7Rα binders were assessed by means of competition BLI, and all six competed for binding with a structurally validated positive control binding to the same site (Supplementary Fig. 10a; further work is required to fully characterize proteome-wide specificity).

We solved the structure of the highest affinity Influenza binder, *HA_20*, in complex with Iowa43 HA using cryo-EM (Extended Data Table 1). Raw electron micrographs revealed a well-folded HA

glycoprotein with clearly discernible side, top and tilted view orientations suspended in a thin layer of vitreous ice (Extended Data Fig. 9a). The 2D class averages further show clear secondary structure elements corresponding to both Iowa43 HA (Extended Data Fig. 9b), as well as the *HA_20* binder bound to the stem (Fig 6e). The 3D heterogenous refinement without symmetry revealed full occupancy of all three HA stem epitopes by the *HA_20* binder. A final non-uniform 3D refinement reconstruction with C3 symmetry yielded a 2.9 Å map of the HA/*HA_20* protein–protein complex (Fig 6f) and corresponding 3D structure that almost perfectly matches the computational design model (0.63 Å, Fig 6f,g; the sidechain interactions at the interface are very different from the closest structure in the PDB; Extended Data Fig. 9h). Over the binder alone, the experimental structure deviates from the RFdiffusion design by only 0.6 Å (Fig. 6h). These results demonstrate the ability of RFdiffusion to generate new proteins with atomic level accuracy, and to precisely target functionally relevant sites on therapeutically important proteins.

Discussion

RFdiffusion is a comprehensive improvement over current protein design methods. RFdiffusion readily generates diverse unconditional designs up to 600 residues in length that are accurately predicted by AF2, far exceeding the complexity and accuracy achieved by most previous methods (a recent Hallucination-based approach also achieved high unconditional performance[53]). Half of our tested unconditional designs express in a soluble way, and have circular dichroism spectra consistent with the design models and high thermostability. Despite their substantially increased complexity, the ideality and stability of RFdiffusion designs is akin to that of de novo protein designs generated using previous methods such as Rosetta. RFdiffusion enables generation of higher-order architectures with any desired symmetry, unlike Hallucination methods, which have so far been limited to cyclic symmetries. Electron microscopy confirmed that the structures of these oligomers are very similar to the design models, which in many cases show little global similarity to known protein oligomers.

There has been recent progress in scaffolding protein functional motifs using deep-learning methods (RF Hallucination, RF_{joint} Inpainting and diffusion), but Hallucination is slow for large systems, Inpainting fails when insufficient starting information is provided and previous diffusion methods had low accuracy. RFdiffusion outperforms these previous methods in the complexity of the motifs that can be scaffolded, the precision with which sidechains are positioned (for catalysis and other functions), and the accuracy of motif recapitulation by AF2. The design of MDM2 binding proteins with three orders of magnitude higher affinities than the scaffolded P53 motif demonstrates the robustness of RFdiffusion motif scaffolding. Combining accurate motif scaffolding with the design of symmetric assemblies enabled consistent and atomically precise positioning of sidechains to coordinate Ni^{2+} ions across diverse tetrameric assemblies

For binder design from target structural information alone, previous work required testing tens of thousands of sequences[12]. RFdiffusion, when combined with improved filtering[26] raises experimental success rates by two orders of magnitude; high-affinity binders can be identified from dozens of designs, in many cases eliminating the requirement for slow and expensive high-throughput screening (at least for the non-polar sites targeted here; further studies will be required to assess success rates on more polar target sites and sites without native binding partners). A high-resolution cryo-EM structure of one of these designs in complex with influenza HA shows that RFdiffusion can design functional proteins with atomic accuracy. Vázquez Torres et al. demonstrate the ability of RFdiffusion to design picomolar affinity binders to flexible helical peptides[54], further highlighting its use for de novo binder design. Vázquez Torres et al. also show how RFdiffusion can be extended for protein model refinement by partial noising and denoising, which enables tuneable sampling around a given input structure. For peptide binder design, this enabled increases in affinity of nearly three orders of magnitude without high-throughput screening.

The breadth and complexity of problems solvable with RFdiffusion and the robustness and accuracy of the solutions far exceeds what has been achieved previously. In a manner reminiscent of the generation of images from text prompts, RFdiffusion makes possible, with minimal specialist knowledge, the generation of functional proteins from minimal molecular specifications (for example, high-affinity binders to a user-specified target protein, and diverse protein assemblies from user-specified symmetries).

The power and scope of RFdiffusion can be extended in several directions. RF has recently been extended to nucleic acids and protein–nucleic acid complexes[55], which should enable RFdiffusion to design nucleic acid binding proteins and perhaps folded RNA structures. Extension of RF to incorporate ligands should similarly enable extension of RFdiffusion to explicitly model ligand atoms, and allow the design of protein–ligand interactions. The ability to customize RFdiffusion to specific design challenges by addition of external potentials and by fine-tuning (as illustrated here for catalytic site scaffolding, binder-targeting and fold specification), along with continued improvements to the underlying methodology, should enable de novo protein design to achieve still higher levels of complexity, to approach and, in some cases, surpass what natural evolution has achieved.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06415-8.

1. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
2. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
3. Singer, J. M. et al. Large-scale design and refinement of stable proteins using sequence-only models. *PLoS ONE* **17**, e0265020 (2022).
4. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
5. Trippe, B. L. et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. in *The Eleventh International Conference on Learning Representations* (2023).
6. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
7. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
8. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. Preprint at https://doi.org/10.48550/arXiv.2205.15019 (2022).
9. Luo, S. et al. Antigen-specific antibody design and optimization with diffusion-based generative models. in *Adv. Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 9754–9767 (Curran Associates, 2022).
10. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. in *Proc. 32nd International Conference on Machine Learning* Vol. 37 (eds Bach, Francis and Blei, David) 2256–2265 (PMLR, 2015).
11. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. in *Adv. Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 6840–6851 (Curran Associates, 2020).
12. Cao, L. et al. Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
13. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
14. Ramesh, A. et al. Zero-shot text-to-image generation. in *Proc. 38th International Conference on Machine Learning* Vol. 139 (eds Meila, M. & Zhang, T.) 8821–8831 (PMLR, 2021).
15. Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. in *Adv. Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 36479–36494 (Curran Associates, 2022).
16. Wu, K. E. et al. Protein structure generation via folding diffusion. Preprint at https://doi.org/10.48550/arXiv.2209.15611 (2022).
17. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

18. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
19. Watson, J. L., Bera, A., Juergens, D., Wang, J. & Baker, D. X-ray crystallographic validation of design from this paper. *Science* **377**, 387–394 (2022).
20. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. Preprint at https://doi.org/10.1101/2022.07.21.500999 (2022).
21. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science* **379**, 1123–1130 (2023).
22. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
23. De Bortoli, V. et al. Riemannian score-based generative modelling. in *Adv. Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 2406–2422 (Curran Associates, 2022).
24. Leach, A., Schmon, S. M., Degiacomi, M. T. & Willcocks, C. G. Denoising diffusion probabilistic models on SO(3) for rotational alignment. In *Proc. ICLR 2022 Workshop on Geometrical and Topological Representation Learning* (2022).
25. Chen, T., Zhang, R. & Hinton, G. Analog bits: generating discrete data using diffusion models with self-conditioning. in *The Eleventh International Conference on Learning Representations* (2023).
26. Bennett, N.R. et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).
27. Anand, N. & Huang, P. Generative modeling for protein structures. in *Adv. Neural Information Processing Systems* Vol. 31 (eds Bengio, S. et al.) (Curran Associates, 2018).
28. Ingraham, J. et al. Illuminating protein space with a programmable generative model. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.01.518682 (2022).
29. Lee, J. S. & Kim, P. M. ProteinSGM: Score-based generative modeling for de novo protein design. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.13.499967 (2022).
30. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
31. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: a de novo protein design framework based on AlphaFold. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.11.463937 (2021).
32. Basanta, B. et al. An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proc. Natl Acad. Sci. USA* **117**, 22135–22145 (2020).
33. Pan, X. et al. Expanding the space of protein geometries by computational design of de novo fold families. *Science* **369**, 1132–1136 (2020).
34. Marcandalli, J. et al. Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. *Cell* **176**, 1420–1431.e17 (2019).
35. Butterfield, G. L. et al. Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
36. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
37. Sterner, R. & Höcker, B. Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055 (2005).
38. Sesterhenn, F. et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**, eaay5051 (2020).
39. Yang, C. et al. Bottom-up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**, 492–500 (2021).
40. Glasgow, A. et al. Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proc. Natl Acad. Sci. USA* **117**, 28046–28055 (2020).
41. Chène, P. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. *Nat. Rev. Cancer* **3**, 102–109 (2003).
42. Kussie, P. H. et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274**, 948–953 (1996).
43. Hunt, A. C. et al. Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice. *Sci. Transl. Med.* **14**, eabn1252 (2022).
44. Silverman, J. et al. Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat. Biotechnol.* **23**, 1556–1561 (2005).
45. Detalle, L. et al. Generation and characterization of ALX-0171, a potent novel therapeutic nanobody for the treatment of respiratory syncytial virus infection. *Antimicrob. Agents Chemother.* **60**, 6–13 (2016).
46. Strauch, E.-M. et al. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat. Biotechnol.* **35**, 667–671 (2017).
47. Boyoglu-Barnum, S. et al. Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021).
48. Walls, A. C. et al. Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382.e17 (2020).
49. Salgado, E. N., Lewis, R. A., Mossin, S., Rheingold, A. L. & Tezcan, F. A. Control of protein oligomerization symmetry by metal coordination: C_2 and C_3 symmetrical assemblies through Cu^{II} and Ni^{II} coordination. *Inorg. Chem.* **48**, 2726–2728 (2009).
50. Salgado, E. N. et al. Metal templated design of protein interfaces. *Proc. Natl Acad. Sci. USA* **107**, 1827–1832 (2010).
51. Quijano-Rubio, A., Ulge, U. Y., Walkey, C. D. & Silva, D.-A. The advent of de novo proteins for cancer immunotherapy. *Curr. Opin. Chem. Biol.* **56**, 119–128 (2020).
52. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
53. Frank, C. et al. Efficient and scalable de novo protein design using a relaxed sequence space. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.24.529906 (2023).
54. Torres, S. V. et al. De novo design of high-affinity protein binders to bioactive helical peptides. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.10.519862 (2022).
55. Baek, M., McHugh, R., Anishchenko, I., Baker, D. & DiMaio, F. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. Preprint at *bioRxiv* https://doi.org/10.1101/2022.09.09.507333 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Design structures, AF2 models and experimental measurements are available at https://figshare.com/s/439fdd59488215753bc3. Cryo-EM maps and corresponding atomic models for the Influenza HA binder in Fig. 6d–h have been deposited in the PDB and the Electron Microscopy Data Bank under accession codes 8SK7 and EMDB-40557, respectively. Electron microscopy data collected for the HE0537 oligomer are available at EMDB-40602.

Code availability

Code for running RFdiffusion has been released on GitHub, free for academic, personal and commercial use at https://github.com/RosettaCommons/RFdiffusion. It is also available as a Google Colab notebook, accessible through GitHub.

56. Yeh, A. H.-W. et al. De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
57. Ribeiro, A. J. M. et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
58. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

Acknowledgements We thank N. Anand and D. Tischer for helpful discussions, and I. Kalvet and Y. Kipnis for providing helpful Rosetta scripts. We thank A. Dosey for the provision of purified influenza HA protein. We thank R. Wu, J. Mou, K. Choi, L. Wu and D. Blei for valuable feedback during writing. We thank I. Haydon for help with graphics. We also thank L. Goldschmidt and K. VanWormer, respectively, for maintaining the computational and wet laboratory resources at the Institute for Protein Design. This work was supported by gifts from Microsoft (D.J., M.B. and D.B.), Amgen (J.L.W.), the Audacious Project at the Institute for Protein Design (B.L.T., I.S., J.Y., H.E. and D.B.), the Washington State General Operating Fund supporting the Institute for Protein Design (P.V. and I.S.), grant no. INV-010680 from the Bill and Melinda Gates Foundation (W.B.A., D.J., J.W. and D.B.), grant no. DE-SC0018940 MOD03 from the US Department of Energy Office of Science (A.J.B. and D.B.), grant no. 5U19AG065156-02 from the National Institute for Aging (S.V.T. and D.B.), an EMBO long-term fellowship no. ALTF 139-2018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R. and D.B.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.R.B.), a Washington Research Foundation Fellowship (S.J.P.), a Human Frontier Science Program Cross Disciplinary Fellowship (grant no. LT000395/2020-C, L.F.M.), an EMBO Non-Stipendiary Fellowship (grant no. ALTF 1047-2019, L.F.M.), the Defense Threat Reduction Agency grant nos. HDTRA1-19-1-0003 (N.H. and D.B.) and HDTRA12210012 (F.D.), the Institute for Protein Design Breakthrough Fund (A.C. and D.B.), an EMBO Postdoctoral Fellowship (grant no. ALTF 292-2022, J.L.W.) and the Howard Hughes Medical Institute (A.C., W.S., R.J.R. and D.B.), an NSF-GRFP (J.Y.), an NSF Expeditions grant (no. 1918839, J.Y., R.B. and T.S.J.), the Machine Learning for Pharmaceutical Discovery and Synthesis consortium (J.Y., R.B. and T.S.J.), the Abdul Latif Jameel Clinic for Machine Learning in Health (J.Y., R.B. and T.S.J.), the DTRA Discovery of Medical Countermeasures Against New and Emerging threats program (J.Y., R.B. and T.S.J.), EPSRC Prosperity Partnership grant no. EP/T005386/1 (E.M.) and the DARPA Accelerated Molecular Discovery program and the Sanofi Computational Antibody Design grant (J.Y., R.B. and T.S.J.). We thank Microsoft and AWS for generous gifts of cloud computing resources.

Author contributions J.L.W., D.J., N.R.B., B.L.T., J.Y. and D.B. conceived the study. J.L.W., D.J., N.R.B., W.A., B.L.T. and J.Y. trained RFdiffusion. B.L.T. and J.Y., with assistance from V.D.B. and E.M., extended diffusion to residue orientations. H.E.E., D.J., J.L.W., N.R.B., N.H., W.S., P.V. and I.S. generated experimentally characterized designs. W.A., B.L.T., J.Y., D.J., J.L.W. and N.R.B. generated computational designs. H.E.E., A.J.B., R.J.R., L.F.M., B.I.M.W., S.J.P., N.H., A.C., S.V.T., J.L.W. and B.L.T. experimentally characterized designs. J.W., A.L. and W.S. contributed additional code. S.O. implemented RFdiffusion on Google Colab. M.B. and F.D. trained RF. D.B., T.S.J. and R.B. offered supervision throughout the project. J.L.W., D.J., B.L.T., N.R.B., J.Y., H.E. and D.B. wrote the manuscript. All authors read and contributed to the manuscript. J.L.W. and D.J. agree that the order of their respective names may be changed for personal pursuits to best suit their own interests.

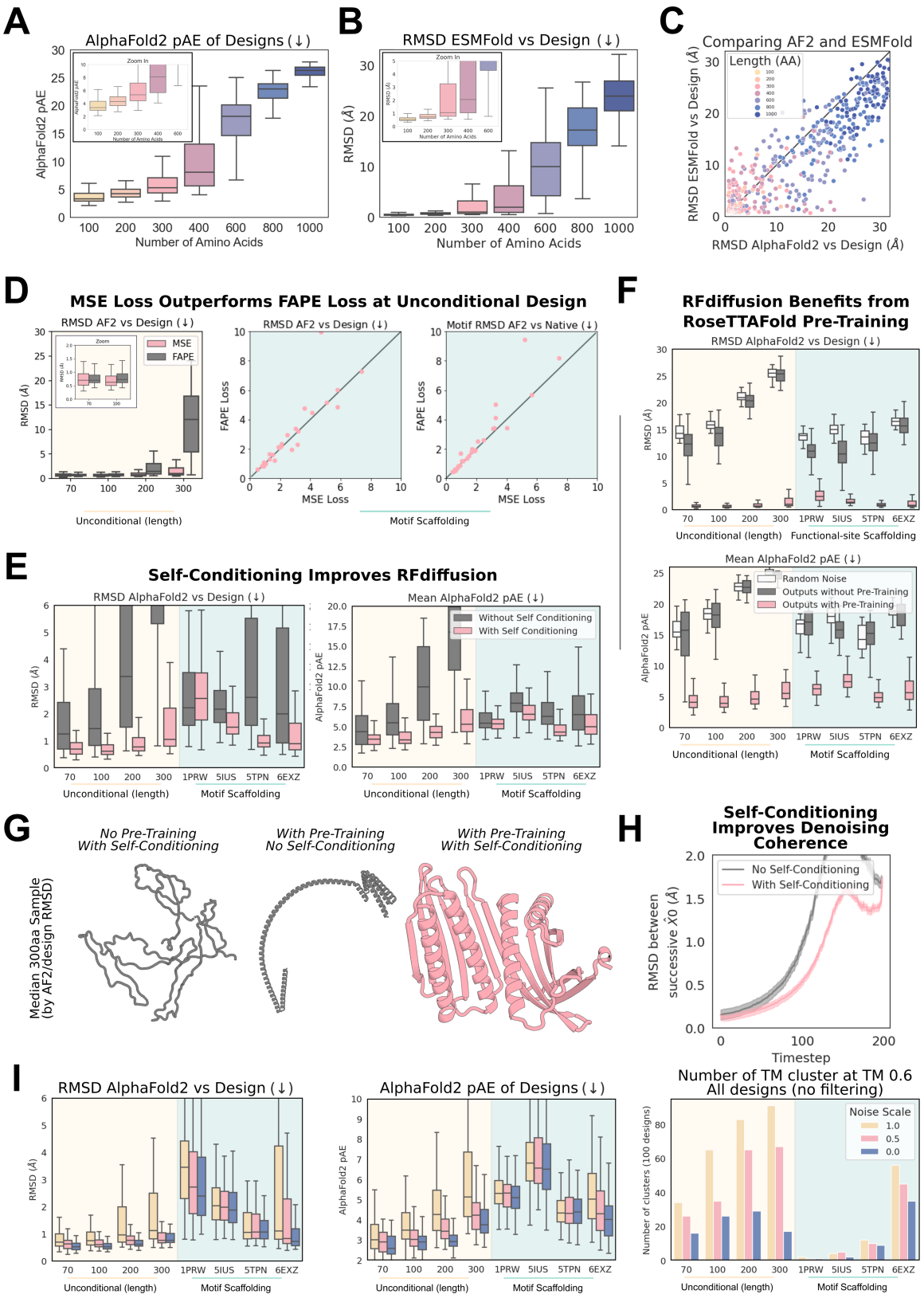
Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06415-8.

Correspondence and requests for materials should be addressed to David Baker.

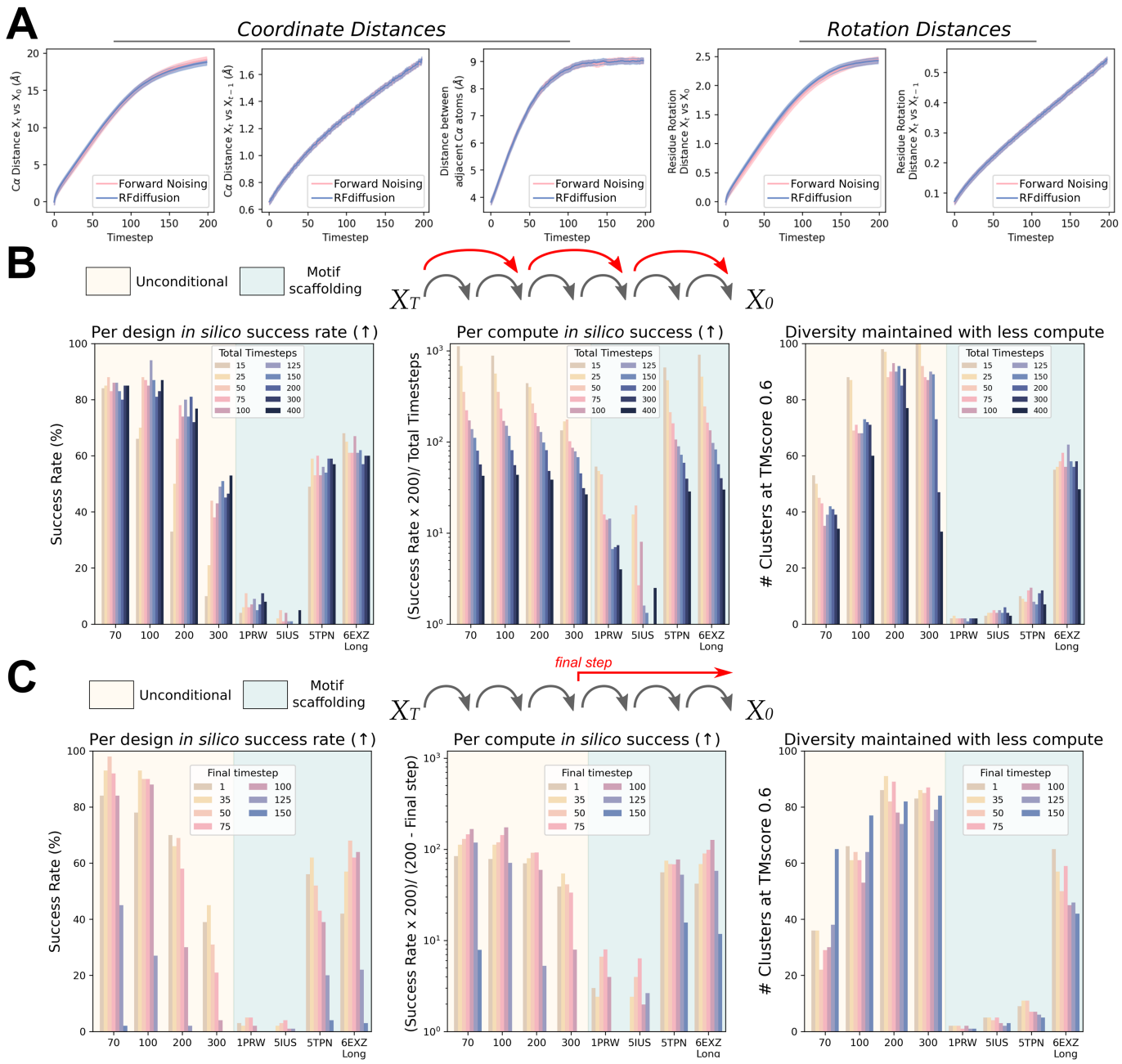
Peer review information *Nature* thanks Arne Elofsson, Giulia Palermo, Alex Pritzel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

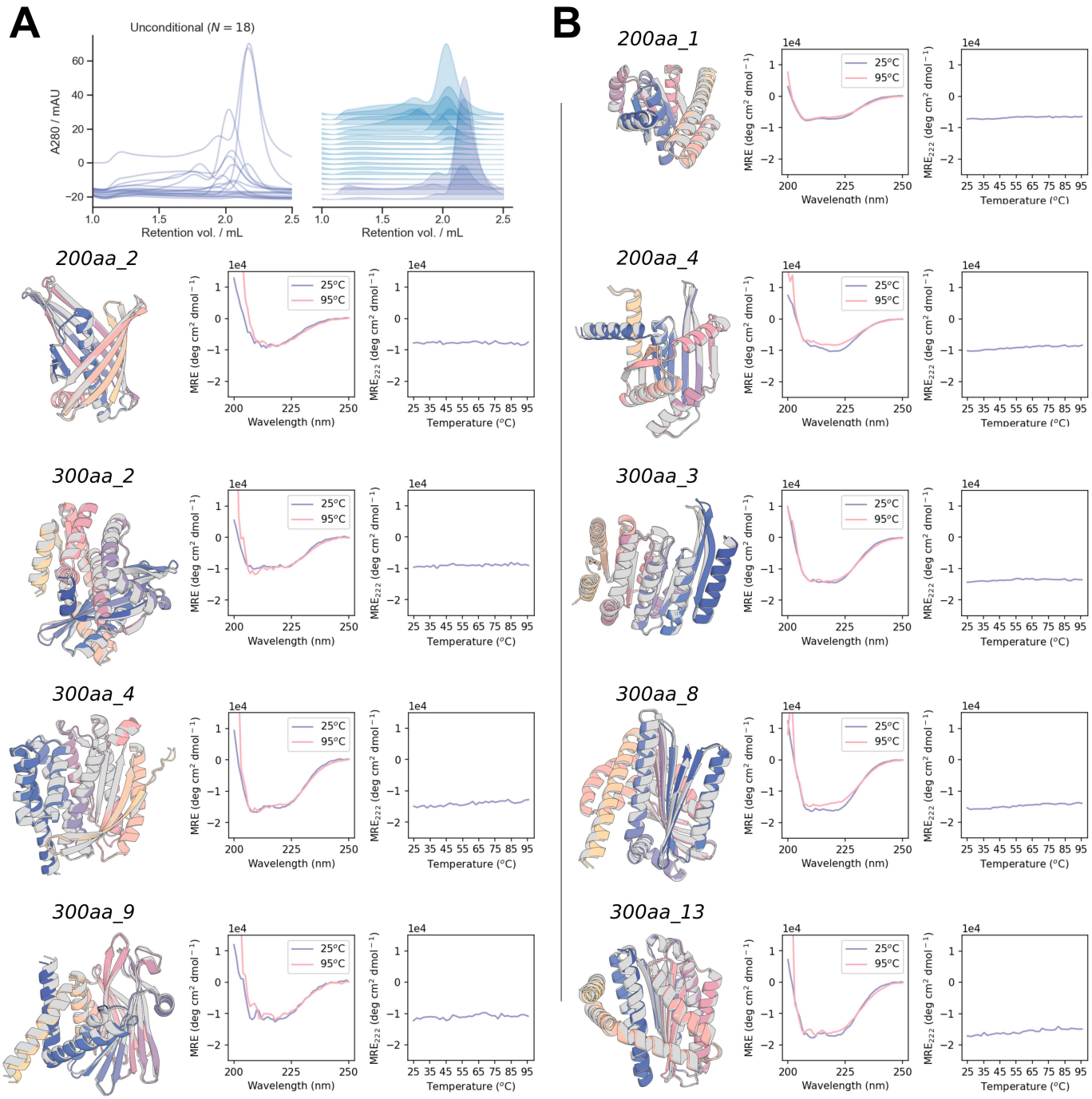


Extended Data Fig. 1 | See next page for caption.

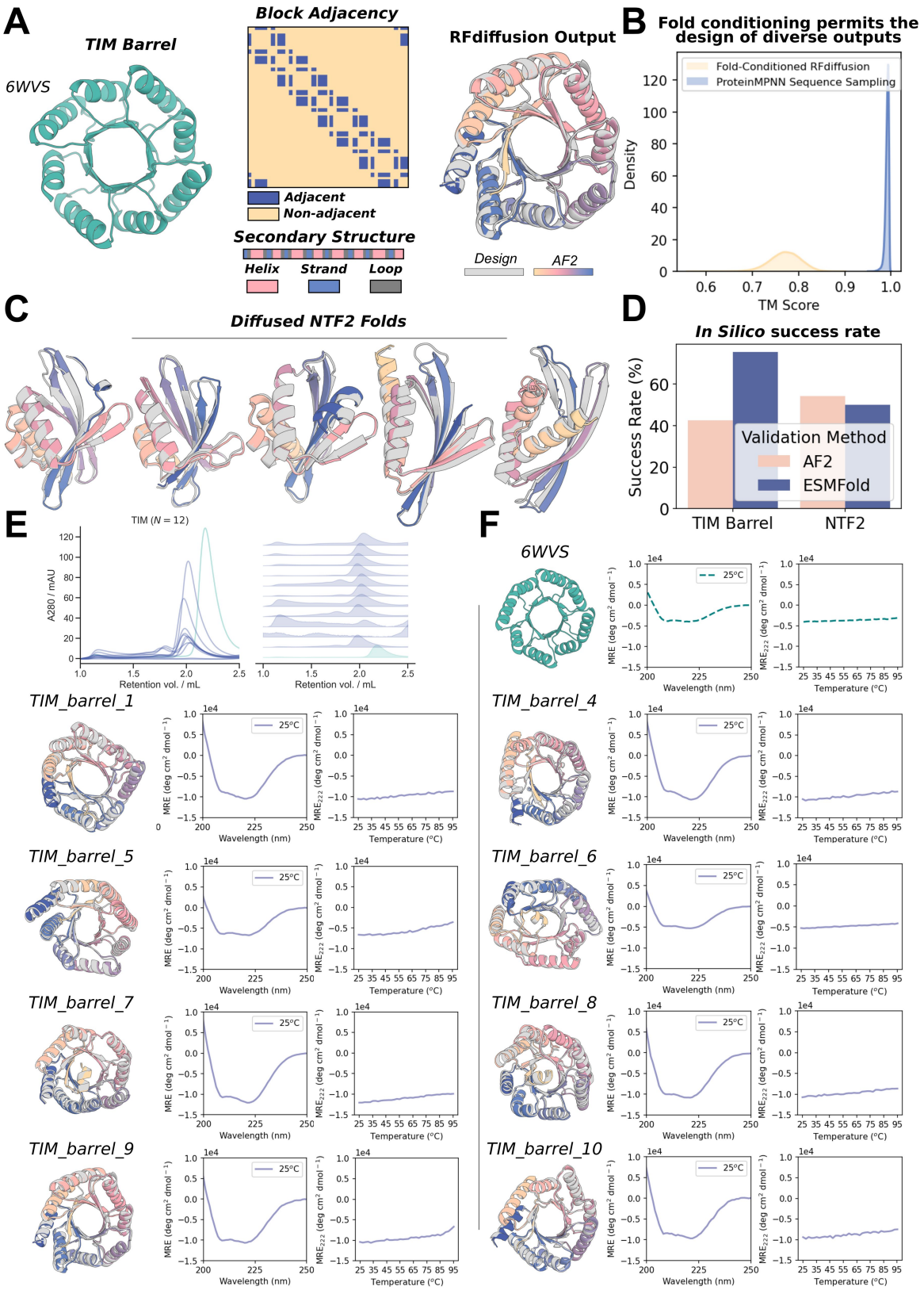
Extended Data Fig. 1 | Training ablations reveal determinants of RFdiffusion success. A–**C**) RFdiffusion can generate high quality large unconditional monomers. Designs are routinely accurately recapitulated by AF2 (see also Fig. 2c), with high confidence (**A**) for proteins up to approximately 400 amino acids in length. **B**) Further orthogonal validation of designs by ESMFold. **C**) Recapitulation of the design structure is often better with ESMFold compared with AF2. For each backbone, the best of 8 ProteinMPNN sequences is plotted, with points therefore paired by backbone rather than sequence. **D**) Comparing RFdiffusion trained with MSE loss on Cα atoms and N-Cα-C backbone frames (Methods 2.5), rather than with FAPE loss[8,17]. The MSE loss is not invariant to the global coordinate frame, unlike FAPE loss, and is required for good performance at unconditional generation (left, two-proportion z-test of in silico success rate, $n = 400$ designs per condition, $z = 4.1, p = 4.1e\text{-}5$). For motif scaffolding problems, where the 'motif' provides a means to align the global coordinate frame between timesteps, FAPE loss performs approximately as well as MSE loss, suggesting the L2 nature of MSE loss (as opposed to the L1 loss in FAPE) is not empirically critical for performance. **E**) Allowing the model to condition on its X_0 prediction at the previous timestep (see Supplementary Methods 2.4) improves designs. Designs with self-conditioning (pink) have improved recapitulation by AF2 (left) and better AF2 confidence in the prediction (right). Two-proportion z-test of in silico success rate, $n = 800$ designs per condition $z = 11.4, p = 6.1e\text{-}30$. **F**) RFdiffusion leverages the representations learned during RF pre-training. RFdiffusion fine-tuned from pre-trained RF (pink) comprehensively outperforms a model trained for an equivalent amount of time, from untrained weights (gray). For context, sequences generated by ProteinMPNN on these output backbones are little better than sampling ProteinMPNN sequences from random Gaussian-sampled coordinates (white). Two-proportion z-test of in silico success rate, pre-training vs without pre-training (or vs random noise; both have zero success rate), $n = 800$ designs per condition, $z = 23.0, p = 3.1e\text{-}117$. Note that the data in pink in **D**–**F** is the same data, reproduced in each plot for clarity. **G**) The median (by AF2 r.m.s.d. vs design) 300 amino acid unconditional sample highlighting the importance of self-conditioning and pre-training. Without pre-training (at least when trained with equivalent compute), RFdiffusion outputs bear little resemblance to proteins (gray, left). Without self-conditioning, outputs show characteristic protein secondary structures, but lack core-packing and ideality (gray, middle). With pre-training and self-conditioning, proteins are diverse and well-packed (pink, right). **H**) Greater coherence during unconditional denoising may partly explain the effect of self-conditioning. Successive X_0 predictions are more similar when the model can self-condition (lower r.m.s.d. between X_0 predictions, pink curve). Data are aggregated from unconditional design trajectories of 100, 200 and 300 residues. **I**) During the reverse (generation) process, the noise added at each step can be scaled (reduced). Reducing the noise scale improves the *in silico* design success rates (left, middle; two-proportion z-test of *in silico* success rate, $n = 800$ designs per condition, 0 vs 0.5: $z = 1.7, p = 0.09$, 0 vs 1: $z = 6.5, p = 6.8e\text{-}11$; 0.5 vs 1: $z = 4.8, p = 1.4e\text{-}6$). This comes at the expense of diversity, with the number of unique clusters at a TM-score cutoff of 0.6 reduced when noise is reduced (right). Note throughout this figure the 6EXZ_long benchmarking problem is abbreviated to 6EXZ for brevity. Boxplots represent median±IQR; tails: min/max excluding outliers (±1.5xIQR).



Extended Data Fig. 2 | RFdiffusion learns the distribution of the denoising process, and inference efficiency can be improved. A) Analysis of simulated forward (noising) and reverse (denoising) trajectories shows that the distribution of Cα coordinates and residue orientations closely match, demonstrating that RFdiffusion has learned the distribution of the denoising process as desired. Left to right: i) average distance between a Cα coordinate at X_t and its position in X_0; ii) average distance between a Cα coordinate at X_t and X_{t-1}; iii) average distance between adjacent Cα coordinates at X_t; iv) average rotation distance between a residue orientation at X_t and X_0; v) average rotation distance between a residue orientation at X_t and X_{t-1}. **B**-**C**) While RFdiffusion is trained to generate samples over 200 timesteps, in many cases, trajectories can be shortened to improve computational efficiency. **B**) Larger steps can be taken between timesteps at inference. Decreasing the number of timesteps speeds up inference, and often does not decrease in silico success rates (left) (for example, on an NVIDIA A4000 GPU, 100 amino acid designs can be generated with 15 steps, in ~11s, with an in silico success rate of over 60%). When normalized for compute budget (center) it is often much more efficient to run more trajectories with fewer timesteps. This can be done without loss of diversity in samples (right). For harder problems (e.g. unconditional 300 amino acids), one must strike an intermediate number of total timesteps (e.g., T = 50) for optimal compute efficiency. Note that for all other analyses in the paper, 200 inference steps were used, in line with how RFdiffusion is trained. **C**) An alternative to taking larger steps is to stop trajectories early (possible because RFdiffusion predicts X_0 at every timestep). In many cases, trajectories can be stopped at timestep 50–75 with little effect on the final in silico success rate of designs (left), and when normalized by compute budget (center), success rates per unit time are typically higher generating more designs with early-stopping. Again, this can be done without a significant loss in diversity (right).

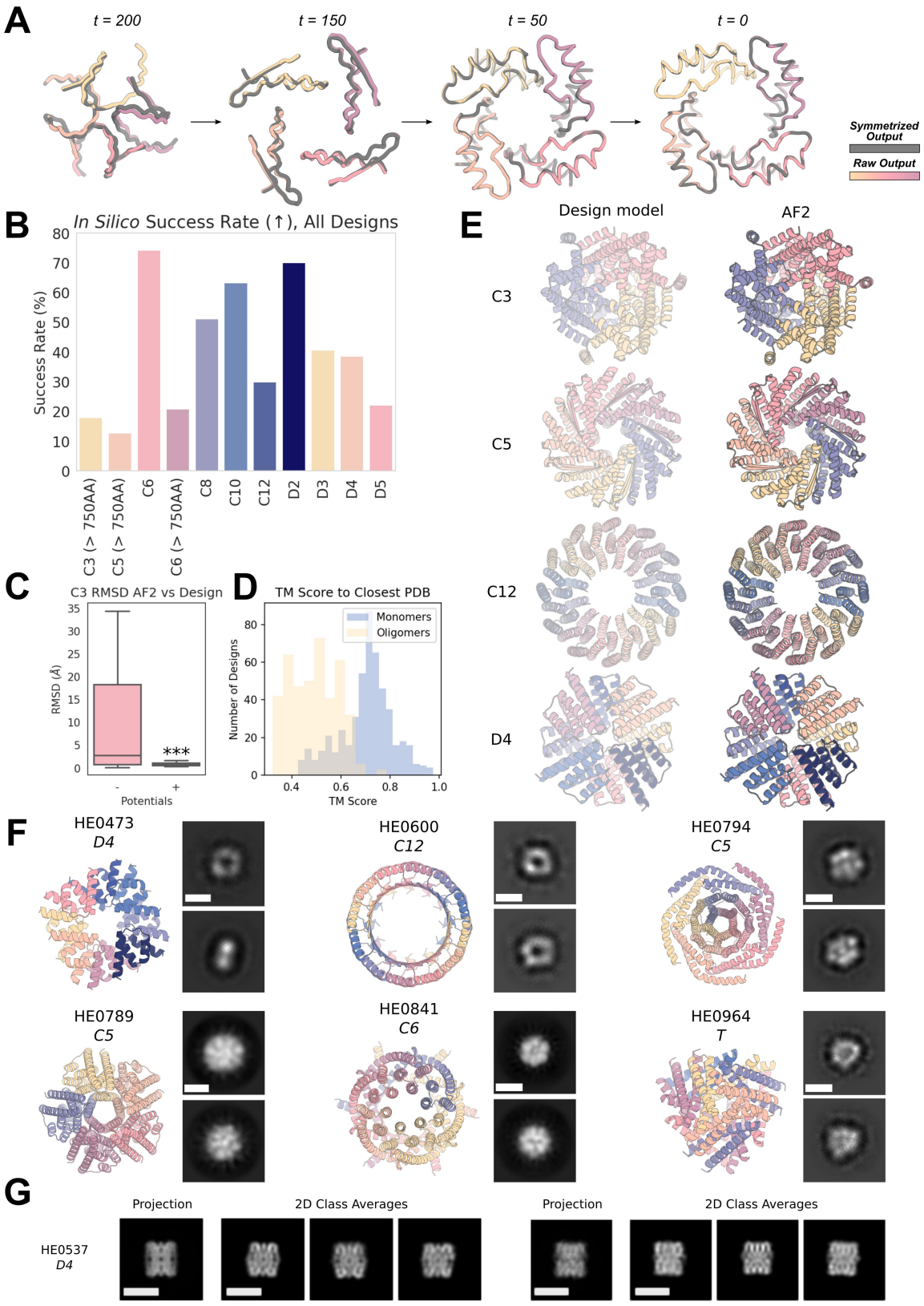


Extended Data Fig. 3 | Unconditionally-generated designs are folded and thermostable. A) Four 200 amino acid and fourteen 300 amino acid proteins were tested for expression and stability. 9/18 designs expressed, with a major peak at the expected elution volume. Blue: 300 amino acid proteins; Purple: 200 amino acid proteins. **B**) Colored AF2 predictions overlaid on gray design models (left), circular dichroism spectra at 25 °C (blue) and 95 °C (pink) (middle) and circular dichroism melt curves (right) for all 9 designs passing expression thresholds. In all cases, proteins remain well folded even at 95 °C. Note that data on *300aa_3* and *300aa_8* are duplicated from Fig. 2f, reproduced here for clarity.



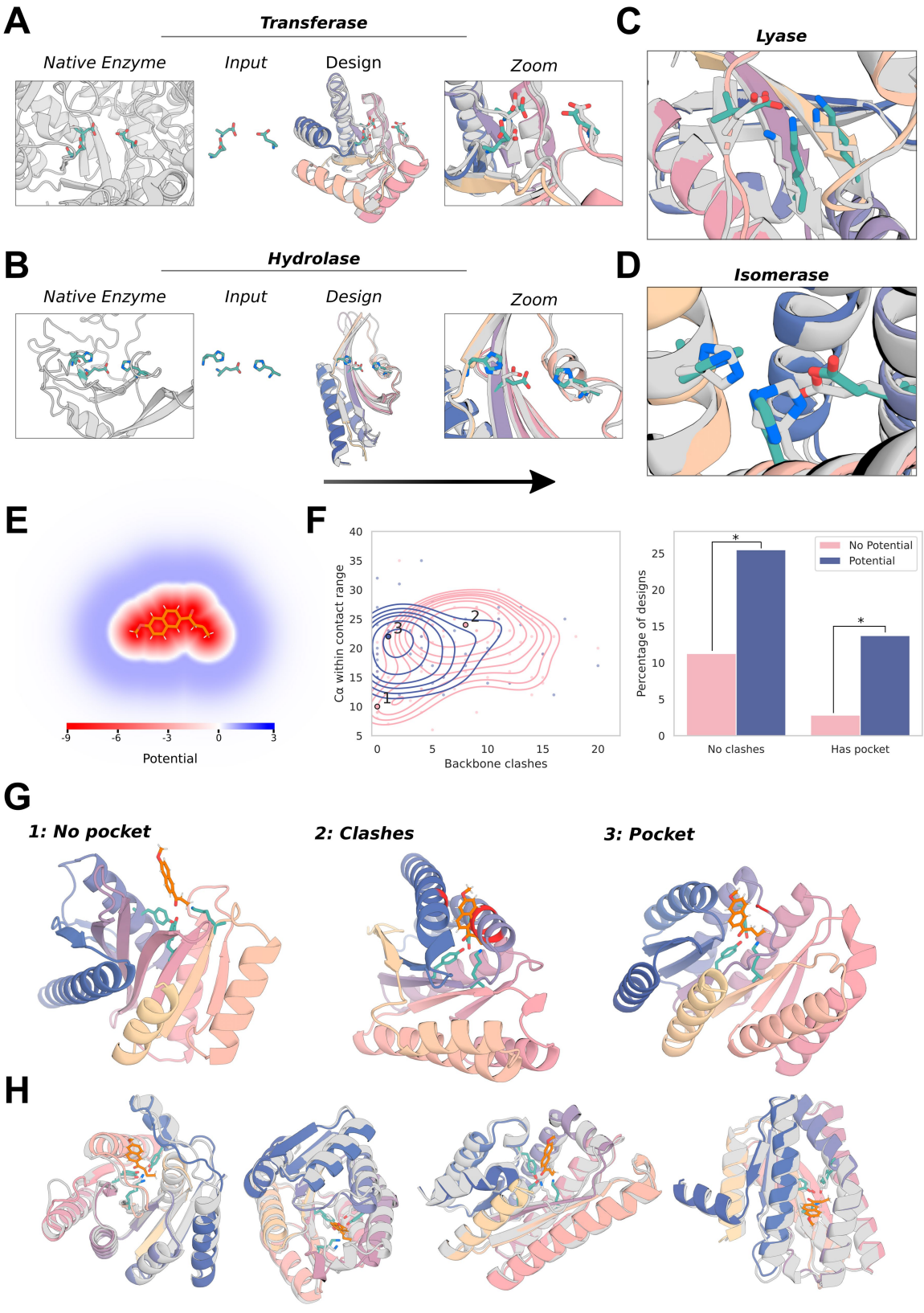
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | RFdiffusion can condition on fold information to generate specific, thermostable folds. A) 6WVS is a previously-described de novo designed TIM barrel (left). A fine-tuned RFdiffusion model can condition on 1D and 2D inputs representing this protein fold, specifically secondary structure (middle, bottom) and block-adjacency information (middle, top) (see Supplementary Methods 4.3.2). RFdiffusion then generates proteins that closely recapitulate this course-grained fold information (right). **B**) Outputs are diverse with respect to each other. With this coarse-grained fold specification, in silico successful designs are much more diverse (as quantified by pairwise TM-scores) compared to diversity generated through simply sampling many sequences for the original PDB backbone (6WVS). **C**) NTF2 folds are useful scaffolds for de novo enzyme design[56], and can also be readily generated with fold-conditioning in RFdiffusion. Designs are diverse and closely recapitulated by AF2. **D**) In silico success rates are high with fold-conditioned diffusion. TIM barrels are generated with an AF2 in silico success rate of 42.5% (left bar, pink) with in silico success incorporating both AF2 metrics and a TM-score vs 6WVS > 0.5. NTF2 folds are generated with an AF2 in silico success rate of 54.1% (right bar, pink), with in silico success incorporating both AF2 metrics and a TM-score vs PDB: 1GY6 > 0.5. In silico success was further validated with ESMFold (blue bars), where a pLDDT > 80 was used as the confidence metric for success. Gray: RFdiffusion design, colors: AF2 prediction. **E**) 11 TIM barrel designs were purified alongside the 6WVS positive control. Ten of these express and elute predominantly as monomers (note that the designs are approximately 4kDa larger than 6WVS). **F**) Eight designs expressed sufficiently for analysis by circular dichroism. All designs are folded, with circular dichroism spectra consistent with the designed structure (middle), and similar to 6WVS. Designs were also all highly thermostable, with CD melt analyses demonstrating designs were folded even at 95 °C (right). Designs are shown in gray, with the AF2 predictions overlaid in colors (left). Note that data on 6WVS and *TIM_barrel_6* are duplicated from Fig. 2g, reproduced here for clarity.



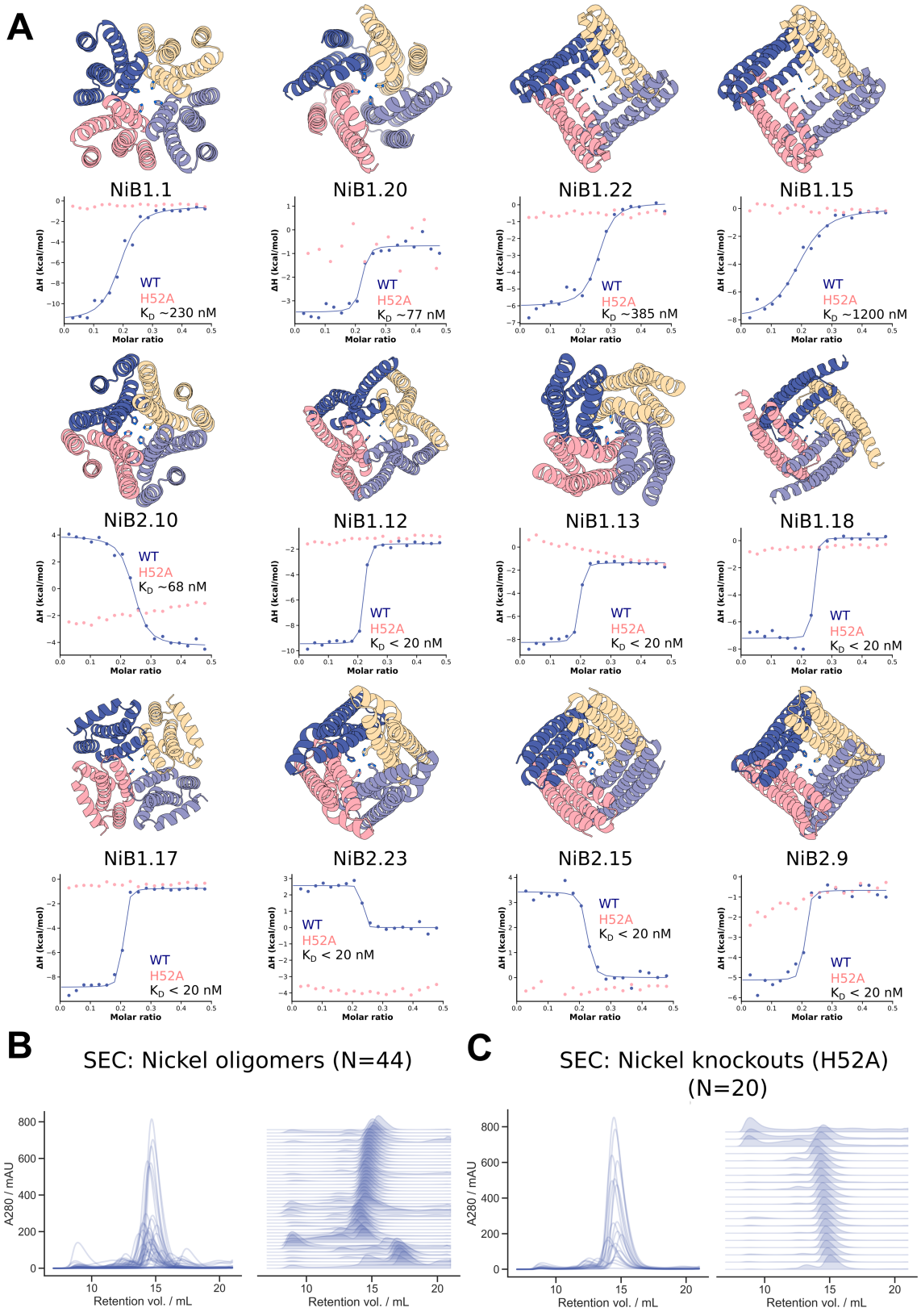
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Symmetric oligomer design with RFdiffusion. A) Due to the (near-perfect - see Supplementary Methods 3.1) equivariance properties of RFdiffusion, X_0 predictions from symmetric inputs are also symmetric, even at very early timepoints (and becoming increasingly symmetric through time; r.m.s.d. vs symmetrized: *t = 200* 1.20 Å; *t = 150* 0.40 Å; *t = 50* 0.06 Å; *t = 0* 0.02Å). Gray: symmetrized (top left) subunit; colors: RFdiffusion X0 prediction. **B**) In silico success rates for symmetric oligomer designs of various cyclic and dihedral symmetries. In silico success is defined here as the proportion of designs for which AF2 yields a prediction from a single sequence that has mean pLDDT > 80 and backbone r.m.s.d. over the oligomer between the design model and AF2 < 2Å. Note that 16 sequences per RFdiffusion design were sampled. **C**) Box plots of the distribution of backbone r.m.s.d.s between AF2 and the RFdiffusion design model with and without the use of external potentials during the trajectory. The external potentials used are the 'inter-chain' contact potential (pushing chains together), as well as the 'intra-chain' contact potential (making chains more globular). Using these potentials dramatically improves in silico success (Two-proportion z-test of in silico success rate: $n = 100$ designs per condition, *z = 4.3, p = 1.9e-5)*. **D**) Designs are diverse with respect to the training dataset (the PDB). While the monomers (typically 60–100 AA) show reasonable alignment to the PDB (median 0.72), the whole oligomeric assemblies showed little resemblance to the PDB (median 0.50). **E**) Additional examples of design models (left) against AF2 predictions (right) for C3, C5, C12, and D4 symmetric designs (the symmetries not displayed in Fig. 3) with backbone r.m.s.d.s (Å) against their AF2 predictions of 0.82, 0.63, 0.79, and 0.78 with total amino acids 750, 900, 960, 640. **F**) Additional nsEM data for symmetric designs. The model is shown on the left and the 2D class averages on the right for each design. **G**) Two orthogonal side views of *HE0537* by cryo-EM. Representative 2D class averages from the cryo-EM data are shown to the right of 2D projection images of the computational design model (lowpass filtered to 8 Å), which appear nearly identical to the experimental data. Scale bars shown (white) are 60 Å. Boxplot represents median ± IQR; tails: min/max excluding outliers (±1.5xIQR).



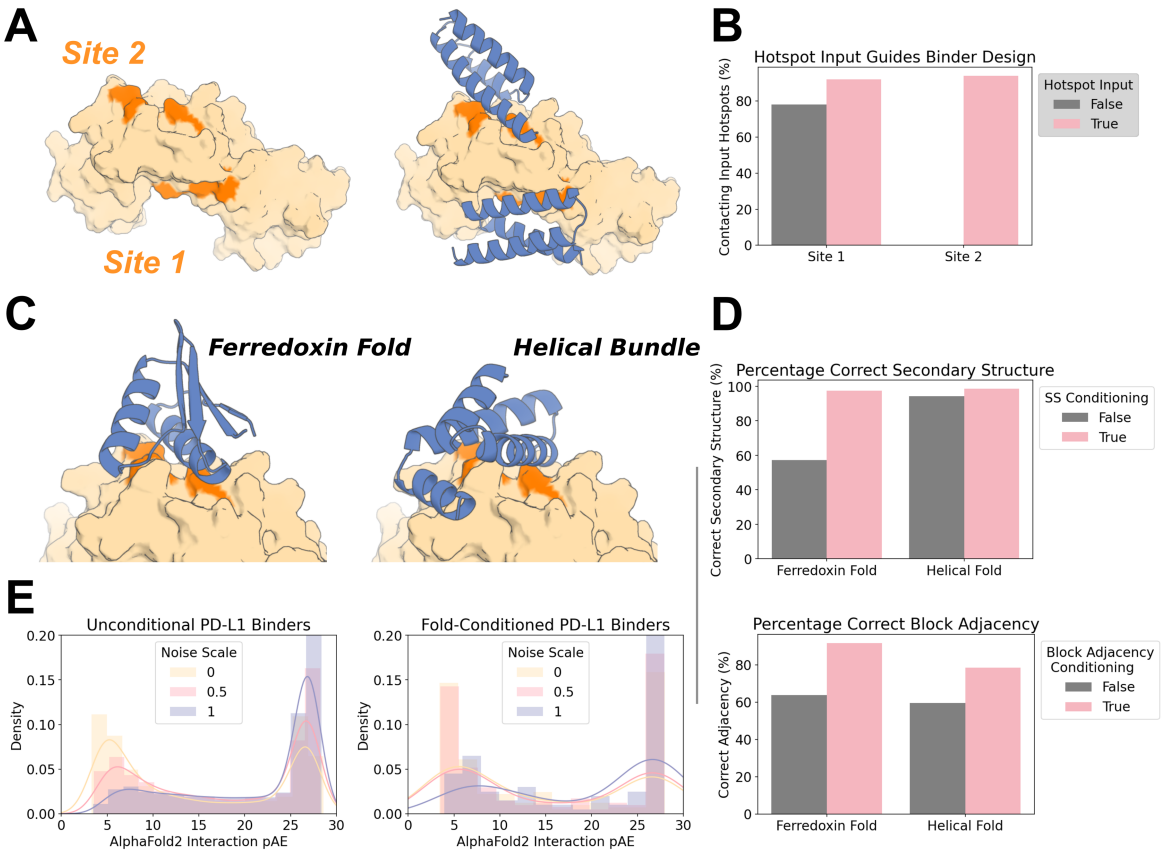
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | External potentials for generating pockets around substrate molecules. A–**D**) Example in silico successful designs for enzyme classes 2–5 (ref. 57, see also Fig. 4). Native enzyme (PDB: 1CWY, 1DE3, 1P1X, 1SNZ); catalytic site (teal); RFdiffusion output (gray: model, colors: AF2 prediction). Metrics (AF2 vs design backbone r.m.s.d., AF2 vs design motif backbone r.m.s.d., AF2 vs design motif full-atom r.m.s.d., AF2 pAE): EC2: 0.93 Å, 0.50 Å, 1.29 Å, 3.51; EC3: 0.92 Å, 0.60 Å, 1.07 Å, 4.59; EC4: 0.93 Å, 0.80 Å, 1.03 Å, 4.41; EC5: 0.78 Å, 0.44 Å, 1.14 Å, 3.32. **E**–**H**) Implicit modeling of a substrate while scaffolding a retroaldolase active site triad [TYR1051-LYS1083-TYR1180] from PDB: 5AN7. **E**) The potential used to implicitly model the substrate, which has both a repulsive and attractive field (see Supplementary Methods 4.4). **F**) Left: Kernel densities demonstrate that without using the external potential (pink), designs often fall into two failure modes: (1) no pocket, and (2) clashes with the substrate. Right: clashes (substrate < 3 Å of the backbone) & pockets (no clash and > 16 Cα within 3–8 Å of substrate) with and without the potential. Two-proportion z-test: $n = 71/51$ +/− potential; clashes $z = -2.05$, $p = 0.02$, pocket $z = -2.27$, $p = 0.01$. Each datapoint represents a design already passing the stringent in silico success metrics (AF2 motif r.m.s.d. < 1 Å, AF2 backbone r.m.s.d. < 2 Å, AF2 pAE < 5). Note that the potential and clash definition pertain only to backbone Cα atoms, and do not currently include sidechain atoms. **G**) Designs close to the labeled local maxima of the kernel density estimate. Without the potential, the catalytic triad is predominantly (*1*) exposed on the surface with no residues available to provide substrate stabilization or (*2*) buried in the protein core, preventing substrate access. With the potential, the catalytic triad is predominantly (*3*), partially buried in a concave pocket with shape complementary to the substrate. Backbone atoms within 3 Å of the substrate are shown in red. **H**) A variety of diverse designs with pockets made using the potential, with no clashes between the substrate and the AF2-predicted backbone. The functional form and parameters used for the pocket potential are detailed in Supplementary Methods 4.4. In each case the substrate is superimposed on the AF2 prediction of the catalytic triad.



Extended Data Fig. 7 | See next page for caption.

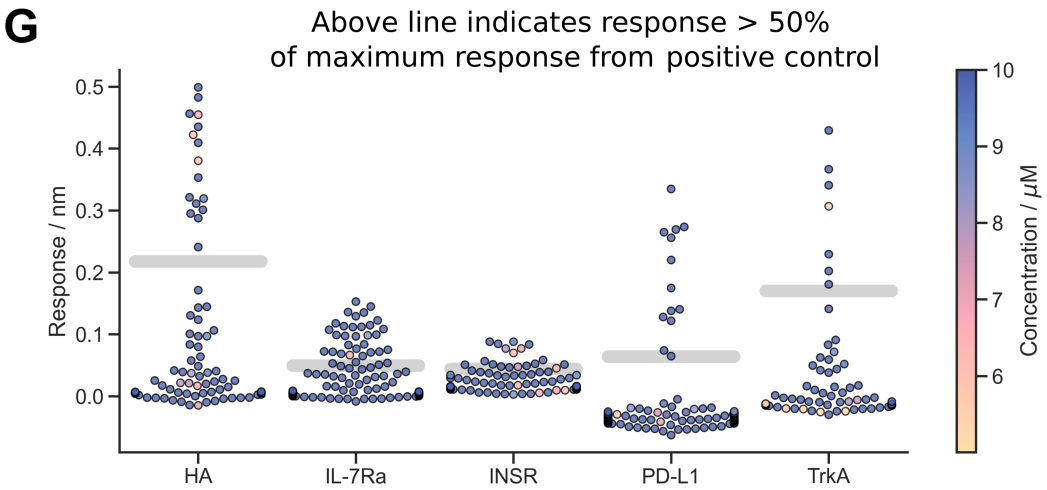
Extended Data Fig. 7 | Additional Ni^{2+} binding C4 oligomers. A) AF2 predictions of a subset of the experimentally verified Ni^{2+} binding oligomers, with corresponding isothermal titration calorimetry (ITC) binding isotherms for the wild-type (blue) and H52A mutant (pink) below. Note that these, with Fig. 5, encompass all of the experimentally validated outputs deriving from unique RFdiffusion backbones. Wild-type dissociation constants are displayed in each plot. We observe a mixture of endothermic (*NiB2.10*, *NiB2.23*, *NiB2.15*) and exothermic isotherms. For all cases displayed we observe no binding to the ion for H52A mutants, indicating the scaffolded histidine at position 52 is critical for ion binding. K_D values in the isotherms indicate binding of the ion with the designed stoichiometry (1:4 Ni^{2+}:protein). Note that each backbone depicted is from a unique RFdiffusion sampling trajectory, and that models and data for designs *NiB2.15*, *NiB1.12*, *NiB1.20* and *NiB1.17* from Fig. 5 are duplicated here for ease of viewing. **B**) Size exclusion chromatograms for elutions from the 44 purifications suggest the vast majority of designs are soluble and have the correct oligomeric state. **C**) Size exclusion chromatograms for 20 H52A mutants show that the mutants remain soluble and retain the intended oligomeric state. Note that only 18 of these 20 had wild-type sequences that definitively bound nickel. Note also that for ITC plots, points represent single measurements.



F

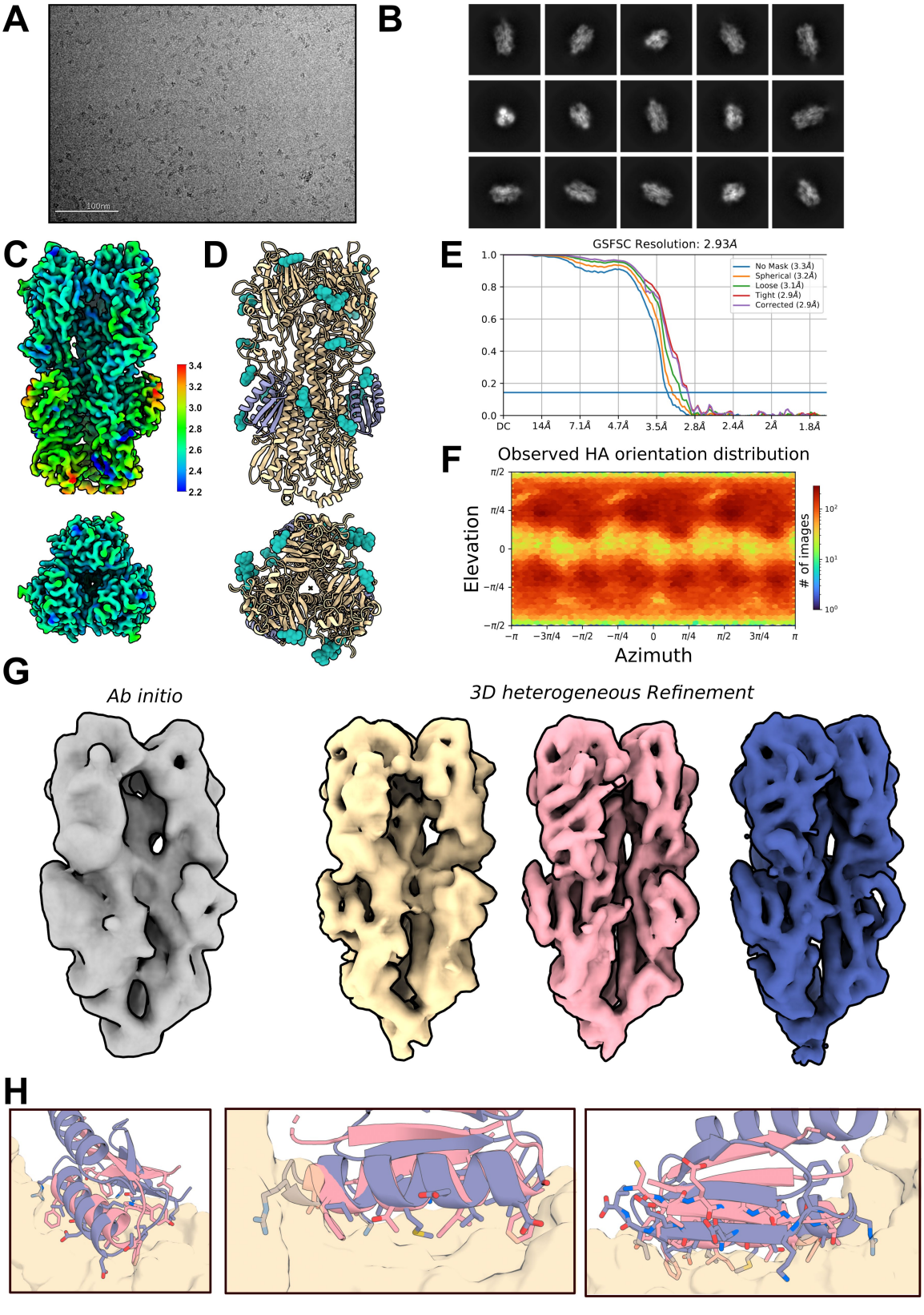
***In Silico* Success Rate (%) [raw / +FastRelax]**

	Noise Scale	PD-L1	IL7 Receptor α	Insulin Receptor	TrkA Receptor	Hemagglutinin
Unconditional	0.0	39.3/47.6	9.1/18.8	**17.9**/17.9	14.7/12.1	4.2/3.9
	0.5	25.4/29.9	6.6/9.9	7.1/6.7	6.5/6.0	0.6/1.0
	1.0	10.7/14.4	2.1/4.4	2.8/3.2	1.5/1.6	0.1/0.1
Fold-Conditioned	0.0	44.0/46.0	11.7/**21.9**	10.6/10.8	**29.2**/24.2	**14.6**/13.9
	0.5	37.0/**49.0**	11.9/15.0	3.7/3.7	16.3/15.2	3.2/4.7
	1.0	22.0/26.0	5.3/10.7	2.5/3.0	6.6/7.2	0.2/2.3



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Targeted unconditional and fold-conditioned protein binder design. A-**B**) The ability to specify where on a target a designed binder should bind is crucial. Specific "hotspot" residues can be input to a fine-tuned RFdiffusion model, and with these inputs, binders almost universally target the correct site. **A**) IL-7Rα (PDB: 3DI3) has two patches that are optimal for binding, denoted Site 1 and Site 2 here. For each site, 100 designs were generated (without fold-specification). **B**) Without guidance, designs typically target Site 1 (left bar, gray), with contact defined as Cα-Cα distance between binder and hotspot reside < 10 Å. Specifying Site 1 hotspot residues increases further the efficiency with which Site 1 is targeted (left bar, pink). In contrast, specifying the Site 2 hotspot residues can completely redirect RFdiffusion, allowing it to efficiently target this site (right bar, pink). **C**-**D**) As well as conditioning on hotspot residue information, a fine-tuned RFdiffusion model can also condition on input fold information (secondary structure and block-adjacency information - see Supplementary Methods 4.5). This effectively allows the specification of a (for instance, particularly compatible) fold that the binder should adopt. **C**) Two examples showing binders can be specified to adopt either a ferredoxin fold (left) or a particular helical bundle fold (right). **D**) Quantification of the efficiency of fold-conditioning. Secondary structure inputs were accurately respected (top, pink). Note that in this design target and target site, RFdiffusion without fold-specification made generally helical designs (right, gray bar). Block-adjacency inputs were also respected for both input folds (bottom, pink). **E**) Reducing the noise added at each step of inference improves the quality of binders designed with RFdiffusion, both with and without fold-conditioning. As an example, the distribution of AF2 interaction pAEs (known to indicate binding when pAE < 10[26]) is shown for binders designed to PD-L1. In both cases, the proportion of designs with interaction pAE < 10 is high (blue curve), and improved when the noise is scaled by a factor 0.5 (pink curve) or 0 (yellow curve). **F**) Full *in silico* success rates for the protein binders designed to five targets. In each case, the best fold-conditioned results are shown (i.e. from the most target-compatible input fold), and the success rates at each noise scale are separated. In line with current best practice[26], we tested using Rosetta FastRelax[58] before designing the sequence with ProteinMPNN, but found that this did not systematically improve designs. *In silico* success is defined in line with current best practice[26]: AF2 pLDDT of the monomer > 80, AF2 interaction pAE < 10, AF2 r.m.s.d. monomer vs design < 1 Å. **G**) Experimentally-validated de novo protein binders were identified for all five of the targets. Designs that bound at 10 μM during single point BLI screening with a response equal to or greater than 50% of the positive control were considered binders. Concentration is denoted by hue for designs that were screened at concentrations less than 10 μM and thus may be false negatives.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Cryo-electron microscopy structure determination of designed Influenza HA binder. A) Representative raw micrograph showing ideal particle distribution and contrast. **B**) 2D Class averages of Influenza H1+*HA_20* binder with clearly defined secondary structure elements and a full-sampling of particle view angles (scale bar = 10 nm). **C**) Cryo-EM local resolution map calculated using an FSC value of 0.143 viewed along two different angles. Local resolution estimates range from ~2.3 Å at the core of H1 to ~3.4 Å along the periphery of the N-terminal helix of the *HA_20* binder. **D**) Cryo-EM structure of the full H1+*HA_20* binder complex (purple: *HA_20*; yellow: H1; teal: glycans). **E**) Global resolution estimation plot. **F**) Orientational distribution plot demonstrating complete angular sampling. **G**) 3D ab initio (left) and 3D heterogenous refinement (right - unsharpened) outputs, performed in the absence of applied symmetry, and showing clear density of the *HA_20* binder bound to all three stem epitopes of the Iowa43 HA glycoprotein trimer, in all maps. **H**) The designed binder has topological similarity to 5VLI, a protein in the PDB, but binds with very different interface contacts.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	HE0537 (EMDB-40602)	H1+HA_20 (EMDB-40557) (PDB 8SK7)
Data collection and processing		
Magnification	36,000	105000
Voltage (kV)	200	300
Electron exposure (e–/Å^2)	65	64.273
Defocus range (μm)	-0.8: -2	0.7-1.8
Pixel size (Å)	0.883	0.84
Symmetry imposed	D4	C3
Initial particle images (no.)	184,703	2,396,954
Final particle images (no.)	36,827	308,846
Map resolution (Å)	6.06	2.93
FSC threshold		
Map resolution range (Å)	5.8-8.47	2.2-3.4
Refinement		
Initial model used (PDB code)		3LZG
Model resolution (Å)		119.6
FSC threshold		
Validation		
MolProbity score		0.92
Clashscore		1.67
Poor rotamers (%)		6
Ramachandran plot		
Favored (%)		98.72
Allowed (%)		1.28
Disallowed (%)		0.00

A.10 Protein function

Recommended reading for Chapter 10 (listed in Section 10.9).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Ashburner et al. (2000) — the Gene Ontology; how function was turned into a computable vocabulary.
  DOI · via SU library (SU login)

  Gene Ontology: tool for the unification of biology. Ashburner et al. (2000). *Nature Genetics*. doi:10.1038/75556

- Whisstock and Lesk (2003) — prediction of protein function from sequence and structure; the classical review.
  DOI · via SU library (SU login)

  Prediction of protein function from protein sequence and structure. Whisstock & Lesk (2003). *Quarterly Reviews of Biophysics*. doi:10.1017/s0033583503003901

  > The sequence of a genome contains the plans of the possible life of an organism, but implementation of genetic information depends on the functions of the proteins and nucleic acids that it encodes. Many individual proteins of known sequence and structure present challenges to the understanding of their function. In particular, a number of genes responsible for diseases have been identified but their specific functions are unknown. Whole-genome sequencing projects are a major source of proteins of unknown function. Annotation of a genome involves assignment of functions to gene products, in most cases on the basis of amino-acid sequence alone. 3D structure can aid the assignment of function, motivating the challenge of structural genomics projects to make structural information available for novel uncharacterized proteins. Structure-based identification of homologues often succeeds where sequence-alone-based methods fail, because in many cases evolution retains the folding pattern long after sequence similarity becomes undetectable. Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function, but are far from foolproof. It is therefore fortunate that the development of whole-organism approaches and comparative genomics permits other approaches to function prediction when the data are available. These include the use of protein-protein interaction patterns, and correlations between occurrences of related proteins in different organisms, as indicators of functional properties. Even if it is possible to ascribe a particular function to a gene product, the protein may have multiple functions. A fundamental problem is that function is in many cases an ill-defined concept. In this article we review the state of the art in function prediction and describe some of the underlying difficulties and successes.

- Marcotte, Pellegrini, Ng, et al. (1999) — the Rosetta stone method; predicting function from genome context rather than homology.
  DOI · via SU library (SU login)

  Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Marcotte et al. (1999). *Science*. doi:10.1126/science.285.5428.751

  > A computational method is proposed for inferring protein interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologs in another organism fused into a single protein chain. Searching sequences from many genomes revealed 6809 such putative protein-protein interactions in Escherichia coli and 45,502 in yeast. Many members of these pairs were confirmed as functionally related; computational filtering further enriches for interactions. Some proteins have links to several other proteins; these coupled links appear to represent functional interactions such as complexes or pathways. Experimentally confirmed interacting pairs are documented in a Database of Interacting Proteins.

- Teufel et al. (2022) — SignalP 6.0; signal peptides of all types predicted with a protein language model.
  full text below · DOI · open access (CC BY 4.0)

  SignalP 6.0 predicts all five types of signal peptides using protein language models. Teufel et al. (2022). *Nature Biotechnology*. doi:10.1038/s41587-021-01156-3

  > Signal peptides (SPs) are short amino acid sequences that control protein secretion and translocation in all living organisms. SPs can be predicted from sequence data, but existing algorithms are unable to detect all known types of SPs. We introduce SignalP 6.0, a machine learning model that detects all five SP types and is applicable to metagenomic data.

- Cheng et al. (2023) — AlphaMissense; proteome-wide prediction of missense variant effects.
  DOI · via SU library (SU login)

  Accurate proteome-wide missense variant effect prediction with AlphaMissense. Cheng et al. (2023). *Science*. doi:10.1126/science.adg7492

  > The vast majority of missense variants observed in the human genome are of unknown clinical significance. We present AlphaMissense, an adaptation of AlphaFold fine-tuned on human and primate variant population frequency databases to predict missense variant pathogenicity. By combining structural context and evolutionary conservation, our model achieves state-of-the-art results across a wide range of genetic and experimental benchmarks, all without explicitly training on such data. The average pathogenicity score of genes is also predictive for their cell essentiality, capable of identifying short essential genes that existing statistical approaches are underpowered to detect. As a resource to the community, we provide a database of predictions for all possible human single amino acid substitutions and classify 89% of missense variants as either likely benign or likely pathogenic.

- Zhou et al. (2019) — CAFA3; how function prediction is evaluated, and what the results show.
  full text below · DOI · open access (CC BY 4.0)

  The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Zhou et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1835-8

  > Background The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function. Results Here, we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in Candida albicans and Pseudomonas aureginosa genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility. We further performed targeted assays on selected genes in Drosophila melanogaster, which we suspected of being involved in long-term memory. Conclusion We conclude that while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than the expectations set by baseline methods in C. albicans and D. melanogaster, it leaves considerable room and need for improvement. Finally, we report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.

The open-access papers are reproduced below under their original licences.

nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-021-01156-3

Check for updates

OPEN

SignalP 6.0 predicts all five types of signal peptides using protein language models

Felix Teufel[1,2], José Juan Almagro Armenteros[3], Alexander Rosenberg Johansen[4], Magnús Halldór Gíslason[5], Silas Irby Pihl[1], Konstantinos D. Tsirigos[6], Ole Winther[5,7,8], Søren Brunak[3], Gunnar von Heijne[9,10] and Henrik Nielsen[1]✉

Signal peptides (SPs) are short amino acid sequences that control protein secretion and translocation in all living organisms. SPs can be predicted from sequence data, but existing algorithms are unable to detect all known types of SPs. We introduce SignalP 6.0, a machine learning model that detects all five SP types and is applicable to metagenomic data.

SPs are short N-terminal amino acid sequences that target proteins to the secretory (Sec) pathway in eukaryotes and for translocation across the plasma (inner) membrane in prokaryotes. As comprehensive experimental identification of SPs is impractical, computational prediction of SPs has high relevance to research in cell biology[1]. SP prediction tools enable identification of proteins that follow the general secretory or twin-arginine translocation (Tat) pathway and predict the position in the sequence where a signal peptidase (SPase) cleaves the SP[2,3]. SignalP 5.0 is able to predict Sec substrates cleaved by SPase I (Sec/SPI) or SPase II (Sec/SPII, prokaryotic lipoproteins) and Tat substrates cleaved by SPase I (Tat/SPI)[4]. However, due to a lack of annotated data, SignalP 5.0 is unable to detect Tat substrates cleaved by SPase II or Sec substrates processed by SPase III (prepilin peptidase, sometimes referred to as SPase IV[2]). Such Sec/SPIII SPs control the translocation of type IV pilin-like proteins, which play a key role in adhesion, motility and DNA uptake in prokaryotes[5]. Furthermore, SignalP 5.0 is agnostic regarding the SP structure, as it cannot define the subregions (the N-terminal n-region, the hydrophobic h-region, and the C-terminal c-region) that underlie the biological function of SPs.

Here, we present SignalP 6.0, based on protein language models (LMs)[6–9] that use information from millions of unannotated protein sequences across all domains of life. LMs create semantic representations of proteins that capture their biological properties and structure. Using these protein representations, SignalP 6.0 can predict additional types of SPs that previous versions have been unable to detect while better extrapolating to both proteins distantly related to those used to create the model and metagenomic data of unknown origin. In addition, it is capable of identifying the subregions of SPs.

We compiled a comprehensive dataset of protein sequences that are known to harbor SPs, containing 3,352 Sec/SPI, 2,261 Sec/SPII, 113 Sec/SPIII, 595 Tat/SPI, 36 Tat/SPII, 16,421 intracellular sequences and 2,615 transmembrane sequences (Methods). Moreover, we defined region-labeling rules according to known properties of the SP types (Fig. 1a and Methods). We applied threefold nested cross-validation to train and evaluate the model (Methods and Supplementary Note 1). In our data-partitioning procedure, we ensured that homologous sequences were placed in the same partition to accurately measure the model's performance on unseen sequences.

For previous predictors, the SP types Sec/SPIII and Tat/SPII were omitted due to a lack of annotated samples, which makes learning their defining features challenging for models[4]. Notably, this lack does not correspond to prevalence in nature, as these types exist throughout most organisms present in the databases[10,11]. In addition, the available annotated sequences do not cover the full diversity encountered in nature, as they are biased towards well-studied organisms. Furthermore, existing predictors require data for which the organism of origin is known, as this allows the predictors to explicitly account for known differences in SP structure among Eukarya, Archaea and Gram-positive and Gram-negative bacteria.

Protein LMs have been shown to improve performance on problems with limited annotated data[12]. Moreover, LM protein representations directly capture the evolutionary context of a sequence[6,8]. We hypothesized that when using an LM, we would (1) obtain better performance on SP types with limited data availability, (2) achieve better generalization to sequences that are distantly related to training sequences and (3) enable the prediction of sequences for which the species of origin is unknown. We opted for the bidirectional encoder representations from transformers (BERT) protein LM, which is available in ProtTrans[6,7] and was trained on UniRef100 (ref. [13]) (Fig. 1b). The LM was subsequently optimized on our dataset to predict SPs. We found that even before optimization, the LM captured the presence of SPs in its protein representations (Fig. 1c). We combined the LM with a conditional random field (CRF) probabilistic model[14] to predict the SP region at each sequence position together with the SP type, yielding the SignalP 6.0 architecture (Fig. 1d).

As the baseline for evaluation, we retrained SignalP 5.0 on our new dataset. We measure performance for each SP type separately per organism group (Archaea, Eukarya, Gram-positive and Gram-negative bacteria), reporting the Matthews correlation coefficient (MCC) for correctly detecting the SP type among both non-SP and other types of sequences as negative samples. For all categories

[1]Section for Bioinformatics, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark. [2]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Center for Genomic Medicine, Rigshospitalet (Copenhagen University Hospital), Copenhagen, Denmark. [6]EMBL-EBI, Wellcome Genome Campus, Cambridge, UK. [7]Department of Biology, Bioinformatics Centre, University of Copenhagen, Copenhagen, Denmark. [8]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark. [9]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden. [10]Science for Life Laboratory, Stockholm University, Solna, Sweden. ✉e-mail: henni@dtu.dk

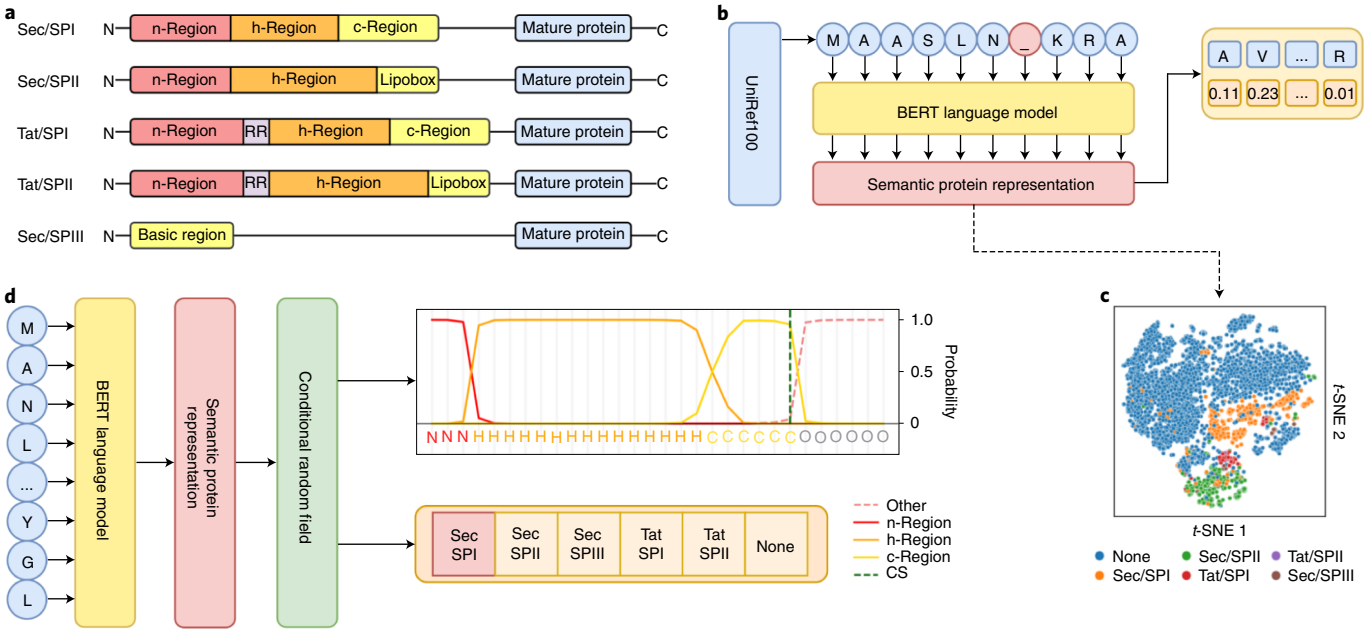


Fig. 1 | Modeling SP structure using protein LMs. a, Region structures of the five SP types. Twin arginine (RR)-translocated SPs feature a twin-arginine motif, while SPs cleaved by SPase II feature a C-terminal lipobox. Sec/SPIII SPs have no substructure. **b**, Protein LM training procedure. BERT learns protein features by predicting masked amino acids in sequences from UniRef100. **c**, *t*-Distributed stochastic neighbor embedding (*t*-SNE) projection of protein representations before prediction training. Different SP types form distinct clusters, separated from sequences without SPs. **d**, SignalP 6.0 architecture. An amino acid sequence is passed through the LM, and the resulting representation serves as input for the CRF, which predicts region probabilities at each position and the SP type. CS, cleavage site.

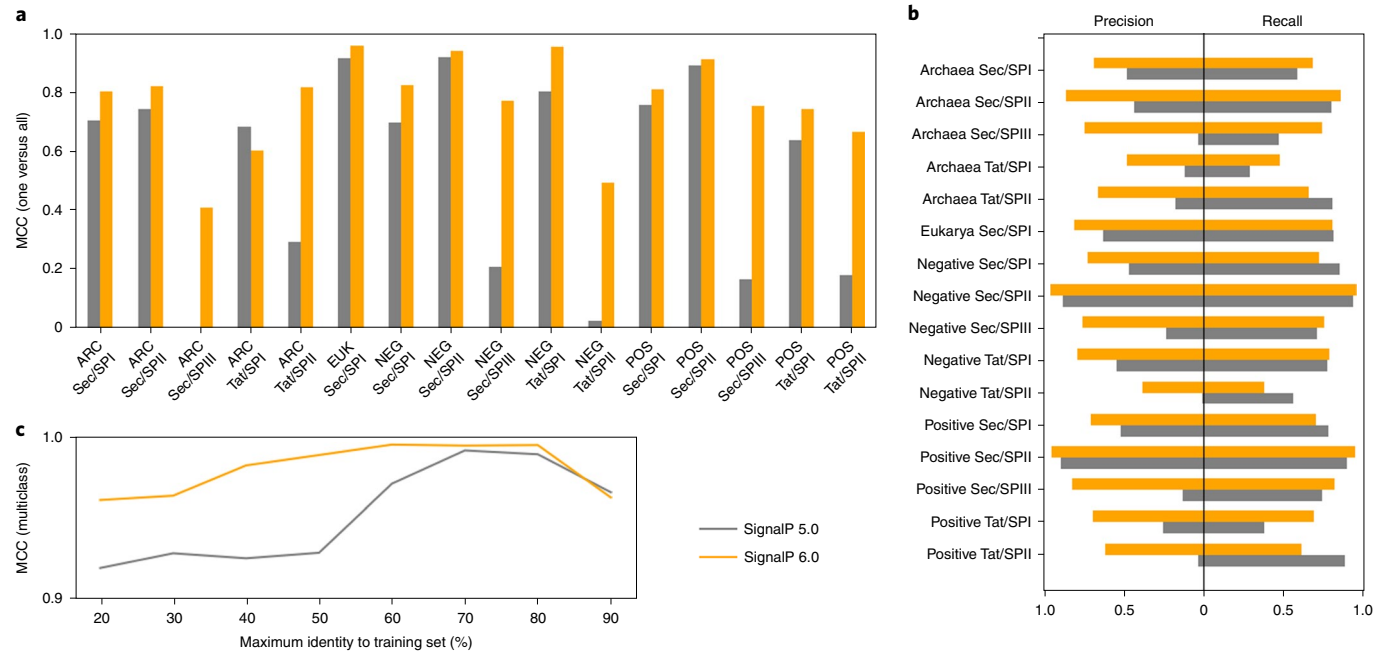


Fig. 2 | SignalP 6.0 shows strong performance on all types and organism groups. a, SP detection performance (ARC, Archaea; EUK, Eukarya; NEG, Gram-negative bacteria; POS, Gram-positive bacteria). SignalP 6.0 substantially improves performance on underrepresented types. **b**, CS prediction performance. SignalP 6.0 has improved precision for all categories. **c**, Dependence of performance on identity to sequences in the training data. At sequence identities lower than 60%, SignalP 6.0 outperforms SignalP 5.0.

except Tat/SPI in Archaea, SignalP 6.0 showed improved performance. Detection performance improved substantially, especially for the two underrepresented types, Sec/SPIII and Tat/SPII (Fig. 2a and Supplementary Fig. 1), whereas the performance of SignalP 5.0 remained too low to make it practically useful. This confirms the importance of LMs for low-data problems, making SignalP 6.0

a model capable of simultaneously detecting all five types of SPs. In addition, we found substantial precision gains for predicting cleavage sites (CSs) (Fig. 2b).

We further benchmarked SignalP 6.0 against other publicly available predictors. In some cases, specialized predictors show stronger performance on the specific tasks they were optimized for (Supplementary Figs. 2 and 3 and Supplementary Tables 1–6). However, none of these predictors are capable of detecting all SP types, and the results are further biased, as they cannot be evaluated in a cross-validated setup.

When predicting a set of test sequences grouped by identity to any sequence in the training data, we find that detection performance at high sequence identities remained comparable. However, at identities lower than 60%, SignalP 6.0 outperformed SignalP 5.0, showing better generalization to proteins distantly related to those present in the training data (Fig. 2c).

Most SP predictors require knowledge of a sequence's organism group of origin for optimal performance[4,15,16]. SignalP 6.0 does not show reduced performance if this information is removed, indicating that the evolutionary context, as encoded in the LM representation, already captures the organism group (Supplementary Fig. 4). Ultimately, this makes SignalP 6.0 a multiclass SP prediction tool that is applicable to sequences of unknown origin, as is typically the case in metagenomic and metatranscriptomic assemblies. However, SignalP 6.0 still relies on start codons being correctly identified before application. For context, 1.7% of UniProt release 2021_02 entries (i.e., 3.5 million sequences) have no organism specified.

SPs are traditionally described as consisting of three regions. We benchmarked our region identification by comparing the properties of predicted regions to known properties[17], with predictions matching all expected properties (Supplementary Note 2 and Supplementary Fig. 5). We additionally predicted a library of synthetic SPs that are either functional or nonfunctional in *Bacillus subtilis*[18], revealing significant differences in the two groups' regions that could not be identified before by traditional sequence analysis (Supplementary Fig. 6 and Supplementary Table 7).

This study presents SignalP 6.0, a machine learning model covering all five known types of SPs that accurately predicts both sequences of unknown origin and evolutionarily distant proteins. Through the use of protein LMs, SignalP 6.0 is able to predict types with very limited training data available. By making the full spectrum of SPs accessible, the model allows us to further improve our understanding of protein translocation throughout evolution (Supplementary Note 3 and Supplementary Tables 8 and 9). In addition, identification of SP regions opens up new avenues into researching the defining properties that govern SP functionality. Given the potential of SPs as drug targets[19] and their emerging role in synthetic biology[18], investigating SPs and their properties at scale may lead to further advances in these fields.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01156-3.

Received: 22 June 2021; Accepted: 8 November 2021;
Published online: 3 January 2022

References

1. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A brief history of protein sorting prediction. *Protein J.* **38**, 200–216 (2019).
2. Dalbey, R. E., Wang, P. & van Dijl, J. M. Membrane proteases in the bacterial protein secretion and quality control pathway. *Microbiol. Mol. Biol. Rev.* **76**, 311–330 (2012).
3. Pohlschroder, M., Pfeiffer, F., Schulze, S. & Halim, M. F. A. Archaeal cell surface biogenesis. *FEMS Microbiol. Rev.* **42**, 694–717 (2018).
4. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
5. Craig, L., Forest, K. T. & Maier, B. Type IV pili: dynamics, biophysics and functional consequences. *Nat. Rev. Microbiol.* **17**, 429–440 (2019).
6. Elnaggar, A. et al. ProtTrans: towards cracking the language of life's code through self-supervised learning. *Trans. Pattern Anal. Mach. Intell.* https://pubmed.ncbi.nlm.nih.gov/34232869/ (2021).
7. Dallago, C. et al. Learned embeddings from deep learning to visualize and predict protein sets. *Curr. Protoc.* **1**, e113 (2021).
8. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
9. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
10. Storf, S. et al. Mutational and bioinformatic analysis of haloarchaeal lipobox-containing proteins. *Archaea* **2010**, 410975 (2010).
11. Hutchings, M. I., Palmer, T., Harrington, D. J. & Sutcliffe, I. C. Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. *Trends Microbiol.* **17**, 13–21 (2009).
12. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
13. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
14. Lafferty, J. D., McCallum, A. & Pereira, F. C. N. In *Proc. 18th International Conference on Machine Learning* (eds. Brodley, C.E. & Danyluk, A.P.) 282–289 (Morgan Kaufmann Publishers, 2001).
15. Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics* **34**, 1690–1696 (2018).
16. Zhang, W.-X., Pan, X. & Shen, H.-B. Signal-3L 3.0: improving signal peptide prediction through combining attention deep learning with window-based scoring. *J. Chem. Inf. Model.* **60**, 3679–3686 (2020).
17. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
18. Wu, Z. et al. Signal peptides generated by attention-based neural networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).
19. Lumangtad, L. A. & Bell, T. W. The signal peptide as a new target for drug design. *Bioorg. Med. Chem. Lett.* **30**, 127115 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
© The Author(s) 2022

Methods

Sequence data. The dataset for SignalP 6.0 was obtained by extending the data published with SignalP 5.0 (ref. [4]). For all classes that were already part of the original data (Sec/SPI, Sec/SPII, Tat/SPI and soluble and transmembrane proteins), we added sequences that had become available in the respective source databases (UniProt[20] and Prosite[21] for SPs and UniProt and TOPDB[22] for soluble and transmembrane proteins) from 2018 until 7 November 2020, following the original selection criteria.

Tat/SPII sequences were identified using the combination of Prosite profiles PS51318 (Tat motif) and PS51257 (lipoprotein motif). By default, PS51318 is subject to postprocessing that prevents both profiles from matching the same sequence. As there is experimental evidence for the existence of Tat-translocated lipoproteins[10,11], we considered this postprocessing rule to be biologically implausible. We disabled it manually in ScanProsite[23] and scanned all prokaryotic sequences in Swiss-Prot, yielding a total of 25 sequences in which both profiles matched. Additional Tat/SPII sequences were found by training a simplified SignalP 6.0 model to discriminate SPII from non-SP sequences. We used this model to predict all Tat/SPI sequences in the training data, as we assumed that PS51257 is not sensitive enough to find all lipoproteins. We investigated the resulting hits in UniProt for supporting evidence that the proteins were true lipoproteins, yielding 12 sequences that we relabeled Tat/SPII from Tat/SPI. One additional sequence with manual evidence was found in the TatLipo 1.03 training data[10]. For Sec/SPIII sequences, we used Prosite pattern PS00409 for bacteria and Pfam[24] family PF04021 for Archaea, yielding 103 and 10 sequences, respectively.

We improved the organism type classification of sequences by defining Gram-negative and Gram-positive bacteria more stringently, as we found that for edge cases such as Thermotogae, in which both gram stains can be observed[25], the classification in SignalP 5.0 was unclear. We redefined Gram positive as all bacterial phyla that have a single membrane (monoderm): Actinobacteria, Firmicutes, Tenericutes, Thermotogae, Chloroflexi and Saccharibacteria. All remaining phyla have a double membrane (diderm) and were classified as Gram negative.

We followed the methodology introduced by Gíslason et al.[26] for homology partitioning of the dataset into three partitions at 30% sequence identity. In brief, it achieves partitioning by computing the pairwise global sequence identities of all sequences using the Needleman–Wunsch algorithm[27], followed by single-linkage clustering. The resulting clusters were grouped together into the desired number of partitions. If there were sequences in a partition that had pairwise identities to any sequence in another partition that were higher than the defined threshold, then those sequences were iteratively removed until the maximum sequence identity criterion was fulfilled. We performed the partitioning procedure separately for each SP type and the negative set, yielding three partitions for each of the six classes. The algorithm was further constrained to ensure that each generated partition was balanced for the four organism groups. We concatenated the resulting 3×6 partitions to yield the three final partitions for cross-validation, thereby ensuring that both the SP types and the organism groups were equally represented across partitions.

The CD-HIT clustering method[28] that was employed in SignalP 5.0 enforces the homology threshold for cluster centers. However, as the training set was not homology reduced but rather homology clustered, other data points can have a homology overlap notably above the chosen threshold of 20% (Supplementary Fig. 7). When using the partitioning method of Gíslason et al., which strictly enforces the defined threshold, 20% maximum identity was impossible to achieve. Even at the relaxed threshold of 30%, the procedure resulted in the removal of a substantial part of the dataset to achieve separation in three partitions (Supplementary Table 10).

For benchmarking, we reused the benchmark set of SignalP 5.0, from which we excluded all sequences that were removed in the homology partitioning procedure of the new dataset. For sequences that were reclassified (to Gram positive or Tat/SPII), we changed the label accordingly (Supplementary Table 11).

For the synthetic SP dataset, we used the data reported by Wu et al.[18]. We gathered all synthetic SP-mature protein pairs that were experimentally characterized, yielding 57 nonfunctional and 52 functional sequences. For the region analysis, we only considered sequences predicted as Sec/SPI SPs by SignalP 6.0, reducing the number of nonfunctional sequences to 55.

Reference proteomes and proteins of unknown origin were obtained from UniProt release 2021_02. To identify sequences of unknown origin, we used taxonomy identifiers 48479 (environmental samples), 49928 (unclassified bacteria) and 2787823 (unclassified entries).

Generation of SP region labels. We defined the task of learning SP regions as a multilabel classification problem at each sequence position. Multilabel differs from multiclass in the sense that more than one label can be true at a given position. This approach was motivated by the fact that there is no strict definition of region borders that is commonly agreed upon, making it impossible to establish ground-truth region labels for models to train on. We thus used the multilabel framework as a method for training with weak supervision, allowing us to use overlapping region labels during the learning phase that could be generated from the sequence data using rules. For inference, we did not make use of the multilabel framework, as we only predicted the single most probable label at each position using Viterbi decoding, yielding a single unambiguous solution.

We defined a set of three rules based on known properties of the n-, h-, and c-regions. The initial n-region must have a minimum length of two residues and the terminal c-region a minimum length of three residues. The most hydrophobic position, which is identified by sliding a seven-amino-acid window across the SP and computing the hydrophobicity using the Kyte–Doolittle scale[29], belongs to the h-region. All positions between these six labeled positions are labeled as either both n and h or h and c, yielding multitag labels.

This procedure was adapted for different SP classes, with only Sec/SPI completely following it. For Tat SPs, the n–h border was identified using the twin-arginine motif. All positions before the motif were labeled n, followed by two dedicated labels for the motif, again followed by a single position labeled n. For SPII SPs, we did not label a c-region, as the C-terminal positions cannot be considered as such[30]. The last three positions were labeled as the lipobox, all positions before that as h only. For SPIII SPs, no region labels were generated within the SP.

Modeling. SignalP 6.0 uses a pretrained protein LM to encode the amino acid sequence and a CRF[14] decoder to predict the regions, CSs and sequence class labels. Specifically, we used the 30-layer BERT LM [31] that is available in ProtTrans[6], which was pretrained on UniRef100 (ref. [13]). We removed the last layer of the pretrained model and extended the pretrained embedding layer by four additional randomly initialized vectors to represent the tokens for the four organism group identifiers. We prepend the organism group identifier to each sequence s of length T and encode it. From the resulting sequence of hidden states, we trim the positions corresponding to the organism group token and the special sequence start and end tokens used by BERT (CLS and SEP) to obtain a sequence of hidden states h of equal length as the original amino acid input x:

$$h = \mathrm{BERT}(x).$$

The hidden states serve as input for a linear-chain CRF. The CRF models the conditional probability of a sequence of states $y = y_1 \ldots y_t$ given a sequence of hidden states $h = \mathbf{h}_1 \ldots \mathbf{h}_t$ using the following factorization:

$$P(y|h) = \frac{1}{Z(h)} \prod_{t=1}^{T} \exp(\psi(\mathbf{h_t})) \prod_{t=1}^{T-1} \exp(\varphi_{y_t, y_{t+1}}),$$

where $Z(h)$ is the normalization constant of the modeled distribution; φ is the learnable transition matrix of the CRF with $C \times C$ parameters, with C being the number of states (labels) modeled by the CRF; and ψ is a learnable linear transformation that maps from the dimension of the hidden state h to the number of CRF states C, yielding the emissions for the CRF:

$$\psi(h_t) = W_\psi \mathbf{h_t} + b_\psi.$$

For each class of SP G, there are multiple possible CRF states, corresponding to the defined regions of the SP class. We constrained the transitions in φ to ensure that regions are predicted in the correct order, leading to the possible state sequences depicted in Supplementary Fig. 8.

For inference, we compute both the most probable state sequence (using Viterbi decoding) and the marginal probabilities at all sequence positions (using the forward-backward algorithm). The most probable state sequence is used to predict the CS, which is inferred from the last predicted SP state as indicated in Supplementary Fig. 8.

As each SP consists of multiple regions, multiple states of C belong to a single global sequence class G. To predict the global class probabilities, we sum the marginal probabilities of all states that belong to a given class and divide the sum by the sequence length. This transforms a matrix of probabilities of shape C × T to a G × 1 vector of global class probabilities:

$$p(G_i|x) = \frac{1}{T} \prod_{t=1}^{T} \sum_{C \in G_i} p(y_{Ct}|x).$$

Training. For training, we minimize the negative log likelihood of the CRF. As we can have multiple true labels y_t at a given position, we use an extension of the equation known as multitag CRF. Multiple labels are handled by summing over the set of true labels M_t at each position:

$$-\log\left(P\left(y|h\right)\right) = \log\left(Z\left(h\right)\right) - \log\left(\exp\left(\sum_{t=1}^{T} \sum_{y_t \in M_t} \psi\left(\mathbf{h_t}\right) + \varphi\left(y_t, y_{t-1}\right)\right)\right).$$

As we designed our region labels to be overlapping, the model is free to distribute its probability mass in any ratio between the correct labels at a given position. There are thus multiple solutions for the specific borders of n-, h- and c-regions that yield the same negative log likelihood but are not equally biologically

plausible. For instance, the model could learn a solution where it uniformly predicts an n-region of length 2 in all SPs, irrespective of the actual sequence. We employ regularization to promote the finding of biologically plausible solutions. Our regularization is based on the fact that the three SP regions have divergent amino acid compositions, which we can quantify by computing the cosine similarity between the amino acid distributions.

The most obvious approach would be to compute the amino acid distribution of each region based on the region borders inferred from the predicted most probable path of the sequence. This, however, cannot be used for regularization, as we require the term to be differentiable, which our Viterbi decoding implementation is not. We therefore based our regularization term on the marginal probabilities of the CRF computed by the forward-backward algorithm, which are used to compute a score for each amino acid for each region, approximating the discrete amino acid distributions.

For each region $r \in \{n, h, c\}$, we sum the marginal probabilities of all CRF states c belonging to region r at position t, yielding $s_{t,r}$. We sum $s_{t,r}$ of all positions t of the sequence that have amino acid a, yielding the elements of the score vector $\mathbf{score}_r$ for each region. We compute the cosine similarity between the normalized score vectors of n and h and h and c:

$$s_{t,r} = \sum_{c \in r} p\left(y_{t,c} | x\right)$$

$$\mathrm{score}_{a,r} = \sum_{t \in I} s_{t,r}$$

$$I = \{t \in T | x_t = a\}$$

$$\mathbf{score_r}' = \mathbf{score_r} / \sum_{a=1}^{A} \mathrm{score}_{a,r}$$

We perform this operation for each sequence. Sequences for which a region does not exist (for example no c-region in Sec/SPII) are ignored for the respective similarity. The mean over all sequences for both similarities, multiplied by a factor α, was added to the loss. We observed that for about half the random seeds we tested, training runs with regularization enabled converged to a n-region length of 2 after one epoch. This is a degenerate solution, as this causes the n-region amino acid distribution to be nonzero at a single position, yielding low similarity scores while being biologically implausible (a length of 2 is expected as the minimum, not the average over all sequences). Such runs were stopped and discarded after one epoch.

The model was trained end-to-end, including all layers of BERT for 15 epochs, using Adamax as the optimizer and a slanted triangular learning rate. We applied dropout on the hidden state outputs of BERT to avoid overfitting. Hyperparameters were optimized using SigOpt (https://app.sigopt.com/docs/intro/overview). We employed threefold nested cross-validation (outer loop is threefold and inner loop is twofold), yielding a total of 3 × 2 models for evaluation.

Evaluation and benchmarking. For comparability, we employed the same metrics that were used in SignalP 5.0. SP detection performance was measured using the MCC[32]. We computed the MCC twice, once with the negative set only consisting of transmembrane and soluble proteins (MCC1) and once with it additionally including sequences of all other SP types (MCC2). Most of the competing single-class predictors considered for benchmarking are optimized for detecting their respective SP type in a dataset of true and non-SP (soluble and transmembrane) sequences; thus, MCC1 best captures their performance on the task they were designed for. MCC2, on the other hand, includes the more challenging task of discriminating between SP types, which is difficult for single-class predictors because of the structural similarity of different SP types. MCC2 represents the performance in most real-world applications, as the presence of a specific SP type usually cannot be ruled out a priori in a set of unknown protein sequences. For CS prediction, we computed the precision and recall. Precision was defined as the fraction of correct CS predictions over the number of predicted CSs, and recall was defined as the fraction of correct CS predictions over the number of true CSs. In both cases, a CS was only considered correct if it was predicted in the correct SP class (e.g., when the model predicts a CS in a Sec/SPI sequence but predicts Sec/SPII as the sequence label, then the sample is considered 'no CS predicted'). To account for possible uncertainty of the CSs in the training data labels, we additionally report these metrics with tolerance windows of one, two and three residues left and right of the true CS (Supplementary Tables 2, 4 and 6).

For the predicted SP regions, in the absence of true labels, no quantitative performance metrics could be established. To still be able to assess the quality of the predictions, we compared the properties of predicted regions with characteristics of regions that are described in the literature. We followed the review by Owji et al.[17] as a guideline to identify region characteristics. Specifically, we evaluated the length, hydrophobicity and charge of each predicted region. Hydrophobicities were computed using the Kyte–Doolittle scale[29], and charges were computed by summing the net charges at pH 7 of all residues. The net charge computation differed between the groups, as in Eukarya and Archaea the N-terminal methionine is not formylated[33], thus contributing an additional positive charge to the n-region by its amino group.

We benchmarked our model against the state-of-the-art model SignalP 5.0, which was reimplemented in PyTorch. Hyperparameter optimization on the new dataset was performed using SigOpt. We also repeated the benchmarking experiment of SignalP 5.0 for all predictors using the adapted benchmark set. We could not add Signal-3L 3.0[16] to the experiment, as the implementation that is available does not allow for processing of more than one sequence at a time, rendering benchmarking intractable. Notably, predictions for all methods except for SignalP 5.0 and SignalP 6.0 were obtained from their publicly available web services, resulting in potential performance overestimation due to the lack of homology partitioning. In addition, performance overestimation is still present for the published version of SignalP 5.0 (named "SignalP 5.0 original" in Supplementary Tables 1–6 and Supplementary Figs. 1 and 2) due to insufficient homology partitioning of its training data by CD-HIT. We thus excluded its values from determining the best-performing tools in the benchmark.

To assess the effect of sequence identity to training sequences on performance, we used the set of sequences that were removed by the partitioning procedure. We predicted all sequences in the removed set and binned the sequences according to the maximum sequence identity to any sequence in the training set. We did this for all six cross-validated models and pooled the resulting binned predictions. For each bin, we computed the multiclass MCC as defined by Gorodkin[34].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article

Data availability

The datasets used for training and testing SignalP 6.0 can be downloaded from https://services.healthtech.dtu.dk/service.php?SignalP-6.0. The investigated reference proteomes are available from UniProt at https://www.uniprot.org/proteomes. The dataset of synthetic SPs was extracted from the supplementary material of the original publication[18] and is included in our GitHub repository.

Code availability

SignalP 6.0 is available at https://services.healthtech.dtu.dk/service.php?SignalP-6.0. The web version of SignalP 6.0 is free for all users, while the standalone Python package is free for academic users (and can be provided upon request) but is licensed for a fee to commercial users. The model source code in PyTorch 1.7 is available at https://github.com/fteufel/signalp-6.0.

References

20. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
21. Sigrist, C. J. A. et al. New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–D347 (2013).
22. Dobson, L., Langó, T., Reményi, I. & Tusnády, G. E. Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.* **43**, D283–D289 (2015).
23. de Castro, E. et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362–W365 (2006).
24. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
25. Bhandari, V. & Gupta, R. S. in *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea* (eds. Rosenberg, E., et al.) 989–1015 (Springer, 2014).
26. Gíslason, M. H., Nielsen, H., Almagro Armenteros, J. J. & Johansen, A. R. Prediction of GPI-anchored proteins with pointer neural networks. *Curr. Res. Biotechnol.* **3**, 6–13 (2021).
27. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
28. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
29. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
30. von Heijne, G. The signal peptide. *J. Membr. Biol.* **115**, 195–201 (1990).
31. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at http://arxiv.org/abs/1810.04805 (2019).
32. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta BBA - Protein Struct.* **405**, 442–451 (1975).
33. Ramesh, V. & RajBhandary, U. L. Importance of the anticodon sequence in the aminoacylation of tRNAs by methionyl-tRNA synthetase and by valyl-tRNA synthetase in an Archaebacterium. *J. Biol. Chem.* **276**, 3660–3665 (2001).
34. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **28**, 367–374 (2004).

Acknowledgements
J.J.A.A. and S.B. were supported by the Novo Nordisk Foundation (grants NNF14CC0001 and NNF17OC0027594). O.W. was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (grant NNF20OC0062606). G.v.H. was supported by the Knut and Alice Wallenberg Foundation (grant 2017.0323), the Novo Nordisk Foundation (grant NNF18OC0032828) and the Swedish Research Council (grant 621-2014-3713).

Author contributions
F.T. collected the datasets, implemented the SignalP 6.0 model and performed the experiments with help from J.J.A.A. and A.R.J. S.I.P. implemented the multitask CRF and developed the Sec/SPI labeling procedure. M.H.G. provided the partitioning code and guided its application. K.D.T., O.W., S.B. and G.v.H. provided suggestions during the design of SignalP 6.0. J.J.A.A., A.R.J. and H.N. supervised and guided the project. All authors edited and approved the manuscript.

Competing interests
The downloadable version of SignalP 6.0 has been commercialized (it is licensed for a fee to commercial users). The revenue from these commercial sales is divided between the program developers and the Technical University of Denmark.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01156-3.

Correspondence and requests for materials should be addressed to Henrik Nielsen.

Peer review information *Nature Biotechnology* thanks Rita Casadio and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Consortium *Genome Biology* (2019) 20:244
https://doi.org/10.1186/s13059-019-1835-8

Genome Biology

RESEARCH Open Access

The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou[1,2], Yuxiang Jiang[3], Timothy R. Bergquist[4], Alexandra J. Lee[5], Balint Z. Kacsoh[6,7], Alex W. Crocker[8], Kimberley A. Lewis[8], George Georghiou[9], Huy N. Nguyen[1,10], Md Nafiz Hamid[1,2], Larry Davis[2], Tunca Dogan[11,37], Volkan Atalay[12], Ahmet S. Rifaioglu[12,13], Alperen Dalkıran[12], Rengul Cetin Atalay[14], Chengxin Zhang[15], Rebecca L. Hurto[16], Peter L. Freddolino[15,16], Yang Zhang[15,16], Prajwal Bhat[17], Fran Supek[18,19], José M. Fernández[20,21], Branislava Gemovic[22], Vladimir R. Perovic[22], Radoslav S. Davidović[22], Neven Sumonja[22], Nevena Veljkovic[22], Ehsaneddin Asgari[23,24], Mohammad R.K. Mofrad[25], Giuseppe Profiti[26,27], Castrense Savojardo[26], Pier Luigi Martelli[26], Rita Casadio[26], Florian Boecker[28], Heiko Schoof[29], Indika Kahanda[30], Natalie Thurlby[31], Alice C. McHardy[32,33], Alexandre Renaux[34,35,36], Rabie Saidi[37], Julian Gough[38], Alex A. Freitas[39], Magdalena Antczak[40], Fabio Fabris[39], Mark N. Wass[40], Jie Hou[41,42], Jianlin Cheng[42], Zheng Wang[43], Alfonso E. Romero[44], Alberto Paccanaro[44], Haixuan Yang[45], Tatyana Goldberg[129], Chenguang Zhao[49], Liisa Holm[50], Petri Törönen[50], Alan J. Medlar[50], Elaine Zosa[51], Itamar Borukhov[51], Ilya Novikov[53], Angela Wilkins[54], Olivier Lichtarge[54], Po-Han Chi[55], Wei-Cheng T seng[56], Michal Linial[57], Peter W. Rose[58], Christophe Dessimoz[59,60,61], Vedrana Vidulin[62], Saso Dzeroski[63,64], Ian Sillit oe[65], Sayoni Das[66], Jonathan Gill Lees[66,67], David T. Jones[69,70], Cen Wan[68,69], Domenico Cozzetto[68,69], Rui Fa[68,69], Mateo Torres[44], Alex Warwick Vesztrocy[70,71], Jose Manuel Rodriguez[72], Michael L. Tress[73], Marco Frasca[74], Marco Notaro[74], Giuliano Grossi[74], Alessandro Petrini[74], Matteo Re[74], Giorgio Valentini[74], Marco Mesiti[74], Daniel B. Roche[76], Jonas Reeb[76], David W. Ritchie[77], Sabeur Aridhi[77], Seyed Ziaeddin Alborzi[77,79], Marie-Dominique Devignes[77,78,79], Da Chen Emily Koo[80], Richard Bonneau[81,82], Vladimir Gligorijević[83], Meet Barot[84], Hai Fang[85], Stefano Toppo[86], Enrico Lavezzo[86], Marco Falda[87], Michele Berselli[86], Silvio C.E. Tosatto[88,89], Marco Carraro[89], Damiano Piovesan[89], Hafeez Ur Rehman[90], Qizhong Mao[91,92], Shanshan Zhang[91], Slobodan Vucetic[91], Gage S. Black[93,94], Dane Jo[93,94], Erica Suh[93], Jonathan B. Dayton[93,94], Dallas J. Larsen[93,94], Ashton R. Omdahl[93,94], Liam J. McGuffin[95], Danielle A. Brackenridge[95], Patricia C. Babbitt[96,98], Jeffrey M. Yunes[97,98], Paolo Fontana[99], Feng Zhang[100,101], Shanfeng Zhu[102,103,104], Ronghui You[102,103,104], Zihan Zhang[102,104], Suyang Dai[102,104], Shuwei Yao[102,103], Weidong Tian[105,106], Renzhi Cao[107], Caleb Chandler[107], Miguel Amezola[107], Devon Jo hnson[107], Jia-Ming Chang[108], Wen-Hung Liao[108], Yi-Wei Liu[108], Stefano Pascarelli[109], Yotam Frank[110], Robert Hoeh ndorf[111], Maxat Kulmanov[111], Imane Boudellioua[112,113], Gianfranco Politano[114], Stefano Di Carlo[114], Alfredo Benso[114], Kai Hakala[115,116], Filip Ginter[115,117], Farrokh Mehryary[115,116], Suwisa Kaewphan[115,116,118], Jari Björne[119,120], Hans Moen[117], Martti E.E. Tolvanen[121], Tapio Salakoski[119,120], Daisuke Kihara[122,123], Aashish Jain[124], Tomislav Šmuc[125], Adrian Altenhoff[126,127], Asa Ben-Hur[128], Burkhard Rost[129,130], Steven E. Brenner[131], Christine A. Orengo[66], Constance J. Jeffery[132], Giovanni Bosco[133], Deborah A. Hogan[6,8], Maria J. Martin[9], Claire O'Donovan[9], Sean D. Mooney[4], Casey S. Greene[134,135], Predrag Radivojac[136*] and Iddo Friedberg[1*]

*Correspondence: predrag@northeastern.edu; idoerg@iastate.edu
[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA
[136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Abstract

Background: The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function.

Results: Here, we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in *Candida albicans* and *aeruginosa* genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility. We further performed targeted assays on selected genes in *Drosophila melanogaster*, which we suspected of being involved in long-term memory.

Conclusion: We conclude that while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than the expectations set by baseline methods in *C. albicans* and *D. melanogaster*, it leaves considerable room and need for improvement. Finally, we report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.

Keywords: Protein function prediction, Long-term memory, Biofilm, Critical assessment, Community challenge

Introduction

High-throughput nucleic acid sequencing [1] and mass-spectrometry proteomics [2] have provided us with a deluge of data for DNA, RNA, and proteins in diverse species. However, extracting detailed functional information from such data remains one of the recalcitrant challenges in the life sciences and biomedicine. Low-throughput biological experiments often provide highly informative empirical data related to various functional aspects of a gene product, but these experiments are limited by time and cost. At the same time, high-throughput experiments, while providing large amounts of data, often provide information that is not specific enough to be useful [3]. For these reasons, it is important to explore computational strategies for transferring functional information from the group of functionally characterized macromolecules to others that have not been studied for particular activities [4–9].

To address the growing gap between high-throughput data and deep biological insight, a variety of computational methods that predict protein function have been developed over the years [10–24]. This explosion in the number of methods is accompanied by the need to understand how well they perform, and what improvements are needed to satisfy the needs of the life sciences community. The Critical Assessment of Functional Annotation (CAFA) is a community challenge that seeks to bridge the gap between the ever-expanding pool of molecular data and the limited resources available to understand protein function [25–27].

The first two CAFA challenges were carried out in 2010–2011 [25] and 2013–2014 [26]. In CAFA1, we adopted a time-delayed evaluation method, where protein sequences that lacked experimentally verified annotations, or *targets*, were released for prediction. After the submission deadline for predictions, a subset of these targets accumulated experimental annotations over time, either as a consequence of new publications about these proteins or the biocuration work updating the annotation databases. The members of this set of proteins were used as *benchmarks* for evaluating the participating computational methods, as the function was revealed only after the prediction deadline.

CAFA2 expanded the challenge founded in CAFA1. The expansion included the number of ontologies used for predictions, the number of target and benchmark proteins, and the introduction of new assessment metrics that mitigate the problems with functional similarity calculation over concept hierarchies such as Gene Ontology [28]. Importantly, we provided evidence that the top-scoring methods in CAFA2 outperformed the top-scoring methods in CAFA1, highlighting that methods participating in CAFA improved over the 3-year period. Much of this improvement came as a consequence of novel methodologies with some effect of the expanded annotation databases [26]. Both CAFA1 and CAFA2 have shown that computational methods designed to perform function prediction outperform a conventional function transfer through sequence similarity [25, 26].

In CAFA3 (2016–2017), we continued with all types of evaluations from the first 2 challenges and additionally performed experimental screens to identify genes associated with specific functions. This allowed us to provide unbiased evaluation of the term-centric performance based on a unique set of benchmarks obtained by assaying *Candida albicans*, *Pseudomonas aeruginosa*, and *Drosophila melanogaster*. We also held a challenge following CAFA3, dubbed CAFA-π, to provide the participating teams another opportunity to develop or modify prediction models. The genome-wide screens on *C. albicans* identified 240 genes previously not known to be involved in biofilm formation, whereas the screens on *P. aeruginosa* identified 532 new genes involved in biofilm formation and 403 genes involved in motility. Finally, we used CAFA predictions to select genes from *D. melanogaster* and assay them for long-term memory involvement. This experiment allowed us to both evaluate prediction methods and identify 11 new fly genes involved in this biological process [29]. Here, we present the outcomes of the CAFA3 challenge, as well as the accompanying challenge CAFA-π, and discuss further directions for the community interested in the function of biological macromolecules.

Results

Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2

One of CAFA's major goals is to quantify the progress in function prediction over time. We therefore conducted a comparative evaluation of top CAFA1, CAFA2, and CAFA3 methods according to their ability to predict Gene Ontology [28] terms on a set of common benchmark proteins. This benchmark set was created as an intersection of CAFA3 benchmarks (proteins that gained experimental annotation after the CAFA3 prediction submission deadline) and CAFA1 and CAFA2 target proteins. Overall, this set contained 377 protein sequences with annotations in the Molecular Function Ontology (MFO), 717 sequences in the Biological Process Ontology (BPO), and 548 sequences in the Cellular Component Ontology (CCO), which allowed for a direct comparison of all methods that have participated in the challenges so far. The head-to-head comparisons in MFO, BPO, and CCO between the top 5 CAFA3 and CAFA2 methods are shown in Fig. 1. CAFA3 and CAFA1 comparisons are shown in Additional file 1: Figure S1.

We first observe that, in effect, the performance of baseline methods [25, 26] has not improved since CAFA2. The Naïve method, which uses the term frequency in the existing annotation database as a prediction score for every input protein, has the same F_{max} performance using both annotation databases in 2014 (when CAFA2 was held) and in 2017 (when CAFA3 was held), which suggests little change in term frequencies in the annotation database since 2014. In MFO, the BLAST method based on the existing annotations in 2017 is slightly but significantly better than the BLAST method based on 2014 training data. In BPO and CCO, however, the BLAST based on the later database has not outperformed its earlier counterpart, although the changes in effect size (absolute change in F_{max}) in both ontologies are small.

When surveying all 3 CAFA challenges, the performance of both baseline methods has been relatively stable, with some fluctuations of BLAST. Such performance of direct sequence-based function transfer is surprising, given the steady growth of annotations in UniProt-GOA [30]; that is, there were 259,785 experimental annotations in 2011, 341,938 in 2014, and 434,973 in 2017, but there does not seem to be a definitive trend with the BLAST method, as they go up and down in F_{max} across ontologies. We conclude from these observations on the baseline methods that first, the ontologies are in different annotation states and should not be treated as a whole. In fact, the distribution of annotation depth and information content is very different across 3 ontologies, as shown in Additional file 1: Figures S15 and S16. Second, methods that perform direct function transfer based on sequence similarity do not necessarily benefit from a larger training dataset. Although the performance observed in our work is also dependent on the benchmark set, it appears that the annotation databases remain too sparsely populated to effectively exploit function transfer by sequence similarity, thus justifying the need for advanced methodology development for this problem.

Head-to-head comparisons of the top 5 CAFA3 methods against the top 5 CAFA2 methods show mixed results. In MFO, the top CAFA3 method, GOLabeler [23], outperformed all CAFA2 methods by a considerable margin, as shown in Fig. 2. The rest of the 4 CAFA3 top methods did not perform as well as the top 2 methods of CAFA2, although only to a limited extent, with little change in F_{max}. Of the top 12 methods ranked in MFO, 7 are from CAFA3, 5 are from CAFA2, and none from CAFA1. Despite the increase in database size, the majority of function prediction methods do not seem to have improved in predicting protein function in MFO since 2014, except for 1 method that stood out. In BPO, the top 3 methods in CAFA3 outperformed their CAFA2 counterparts, but with very small margins. Out of the top 12 methods in BPO, 8 are from CAFA3, 4 are from CAFA2, and none from CAFA1. Finally, in CCO, although 8 out of the top 12 methods over all CAFA challenges come from CAFA3, the top method is from CAFA2. The differences between the top-performing methods are small, as in the case of BPO.

The performance of the top methods in CAFA2 was significantly better than of those in CAFA1, and it is interesting to note that this trend has not continued in CAFA3.

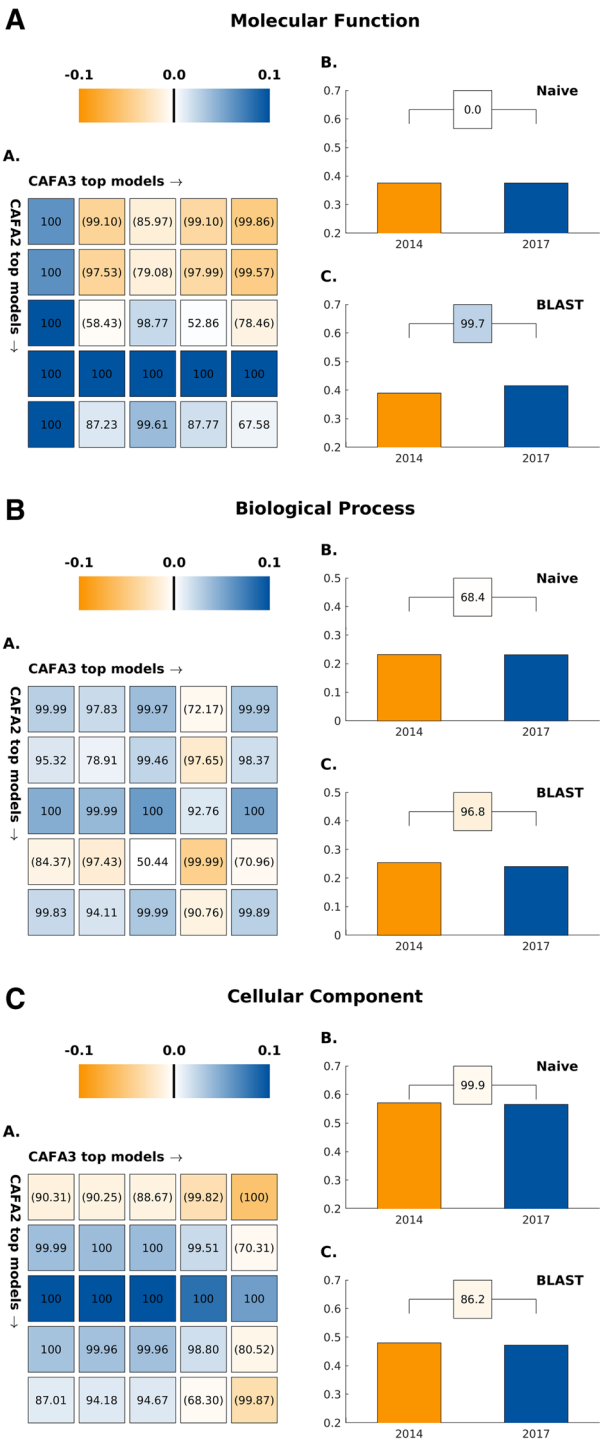


Fig. 1 A comparison in F_{max} between the top 5 CAFA2 models against the top 5 CAFA3 models. Colored boxes encode the results such that (1) the colors indicate margins of a CAFA3 method over a CAFA2 method in F_{max} and (2) the numbers in the box indicate the percentage of wins. **a** CAFA2 top 5 models (rows, from top to bottom) against CAFA3 top 5 models (columns, from left to right). **b** Comparison of the performance (F_{max}) of Naïve baselines trained respectively on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. **c** Comparison of the performance (F_{max}) of BLAST baselines trained on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. Statistical significance was assessed using 10,000 bootstrap samples of benchmark proteins

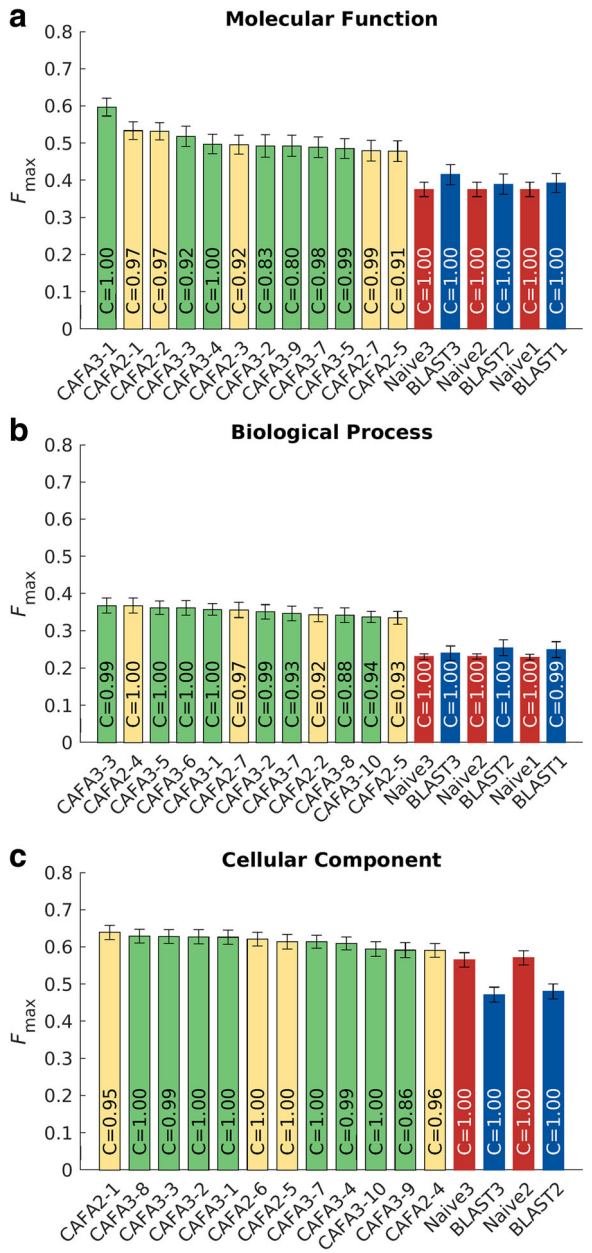


Fig. 2 Performance evaluation based on the F_{max} for the top CAFA1, CAFA2, and CAFA3 methods. The top 12 methods are shown in this barplot ranked in descending order from left to right. The baseline methods are appended to the right; they were trained on training data from 2017, 2014, and 2011, respectively. Coverage of the methods were shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark that are predicted by the methods. Color scheme: CAFA2, ivory; CAFA3, green; Naïve, red; BLAST, blue. Note that in MFO and BPO, CAFA1 methods were ranked, but since none made to the top 12 of all 3 CAFA challenges, they were not displayed. The CAFA1 challenge did not collect predictions for CCO. **a**: molecular function; **b**: Biological process; **c**: Cellular Component

This could be due to many reasons, such as the quality of the benchmark sets, the overall quality of the annotation database, the quality of ontologies, or a relatively short period of time between challenges.

Protein-centric evaluation

The *protein-centric* evaluation measures the accuracy of assigning GO terms to a protein. This performance is shown in Figs. 3 and 4.

We observe that all top methods outperform the baselines with the patterns of performance consistent with CAFA1 and CAFA2 findings. Predictions of MFO terms achieved the highest F_{max} compared with predictions in the other two ontologies. BLAST outperforms Naïve in predictions in MFO, but not in BPO or CCO. This is because sequence similarity-based methods such as BLAST tend to perform best when transferring basic biochemical annotations such as enzymatic activity. Functions in biological process, such as pathways, may not be as preserved by sequence similarity, hence the poor BLAST performance in BPO. The reasons behind the difference among the three ontologies include the structure and complexity of the ontology as well as the state of the annotation database, as discussed previously [26, 31]. It is less clear why the performance in CCO is weak, although it might be hypothesized that such performance is related to the structure of the ontology itself [31].

The top-performing method in MFO did not have as high an advantage over others when evaluated using the S_{min} metric. The S_{min} metric weights GO terms by conditional information content, since the prediction of more informative terms is more desirable than less informative, more general, terms. This could potentially explain the smaller gap between the top predictor and the rest of the pack in S_{min}. The weighted F_{max} and normalized S_{min} evaluations can be found in Additional file 1: Figures S4 and S5.

Species-specific categories

The benchmarks in each species were evaluated individually as long as there were at least 15 proteins per species. Here, we present the results from eukaryotic and bacterial species (Fig. 5). We observed that different methods could perform differently on different species. As shown in Fig. 6, bacterial proteins make up a small portion of all benchmark sequences, so their effects on the performances of the methods are often masked. Species-specific analyses are therefore useful to researchers studying certain organisms. Evaluation results on individual species including human (Additional file 1: Figure S6), *Arabidopsis thaliana* (Additional file 1: Figure S7) and *Escherichia coli* (Additional file 1: Figure S10) can be found in Additional file 1: Figure S6-S14.

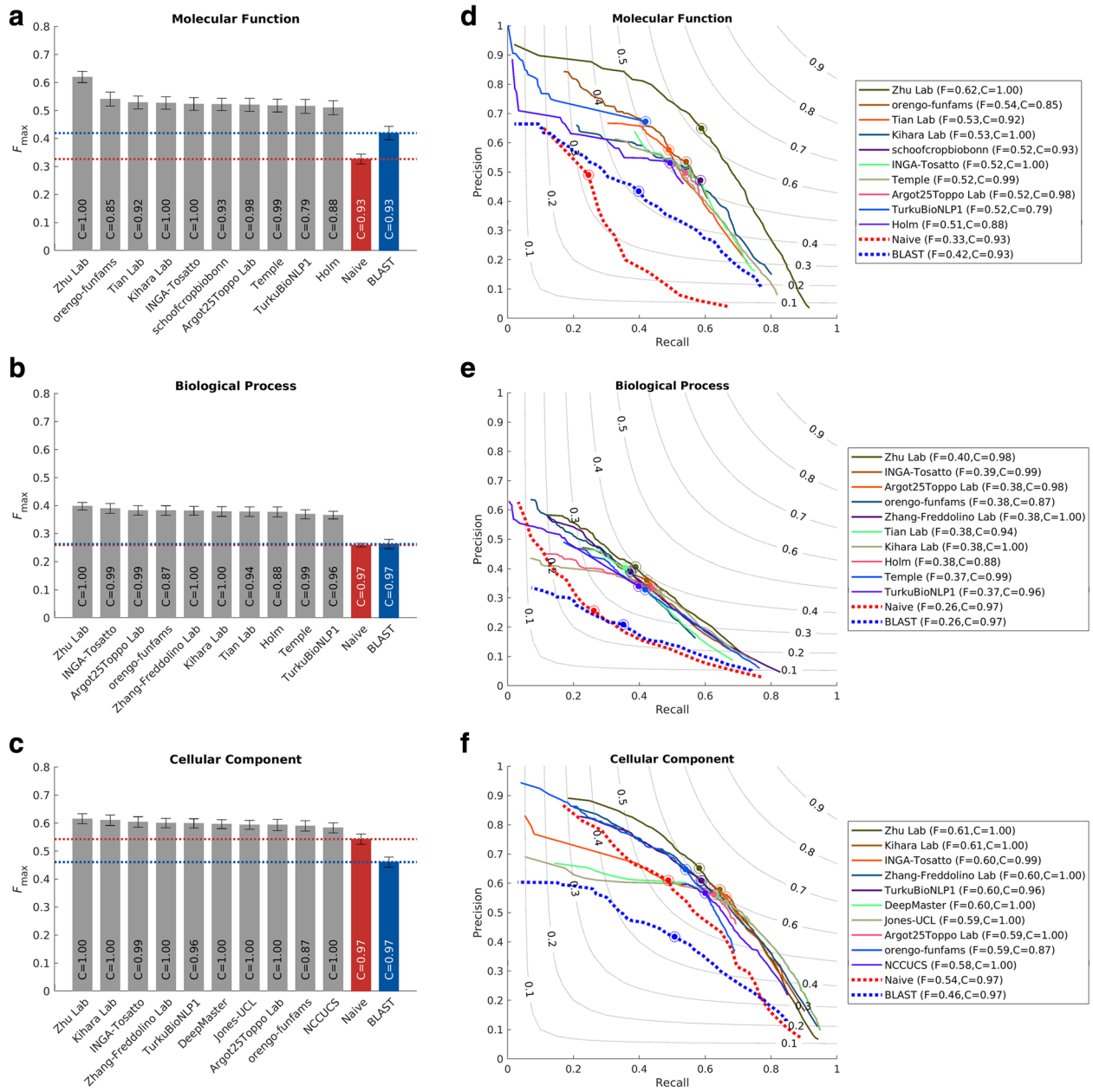


Fig. 3 Performance evaluation based on the F_{max} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing the F_{max} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: precision-recall curves for the top 10 methods. The perfect prediction should have $F_{max} = 1$, at the top right corner of the plot. The dot on the curve indicates where the maximum *F* score is achieved

Diversity of methods

It was suggested in the analysis of CAFA2 that ensemble methods that integrate data from different sources have the potential of improving prediction accuracy [32]. Multiple data sources, including sequence, structure, expression profile, genomic context, and molecular interaction data, are all potentially predictive of the function of the protein. Therefore, methods that take advantage of these rich sources as well as existing techniques from other research groups might see improved performance. Indeed, the one method that stood out from the rest in CAFA3 and performed significantly better than all methods across three challenges is a machine learning-based ensemble method [23]. Therefore, it is important to analyze what information sources and prediction algorithms are better at predicting function. Moreover, the similarity

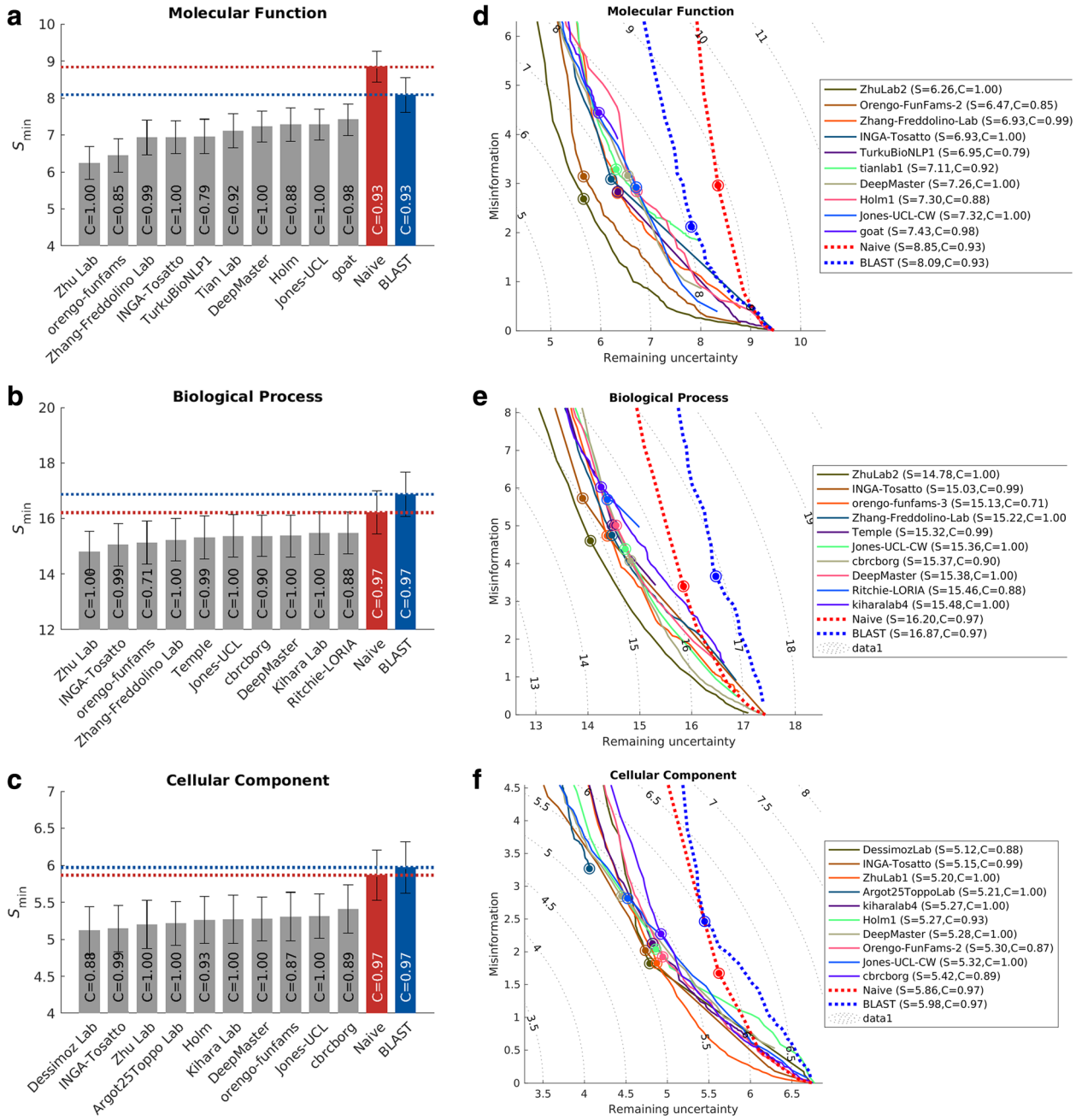


Fig. 4 Performance evaluation based on S_{min} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing S_{min} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: remaining uncertainty-missing information (RU-MI) curves for the top 10 methods. The perfect prediction should have $S_{min} = 0$, at the bottom left corner of the plot. The dot on the curve indicates where the minimum semantic distance is achieved

of the methods might explain the limited improvement in the rest of the methods in CAFA3.

The top CAFA2 and CAFA3 methods are very similar in performance, but that could be a result of aggregating predictions of different proteins to one metric. When computing the similarity of each pair of methods as the Euclidean distance of prediction scores (Fig. 7), we are not interested whether these predictions are correct according to the benchmarks, but simply whether they are similar to one another. The diagonal blocks in Fig. 7 show that

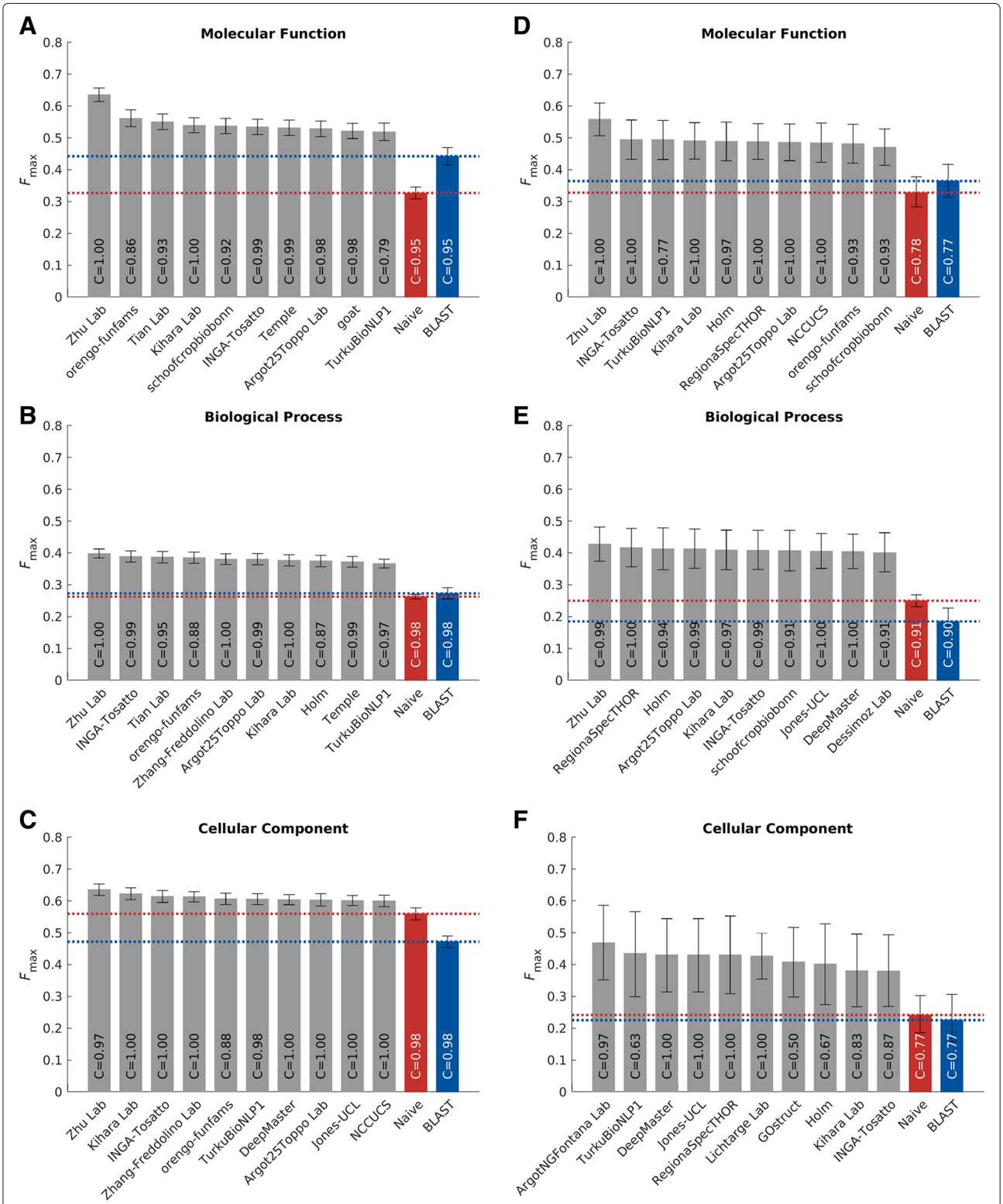


Fig. 5 Evaluation based on the F_{max} for the top-performing methods in eukaryotic and bacterial species

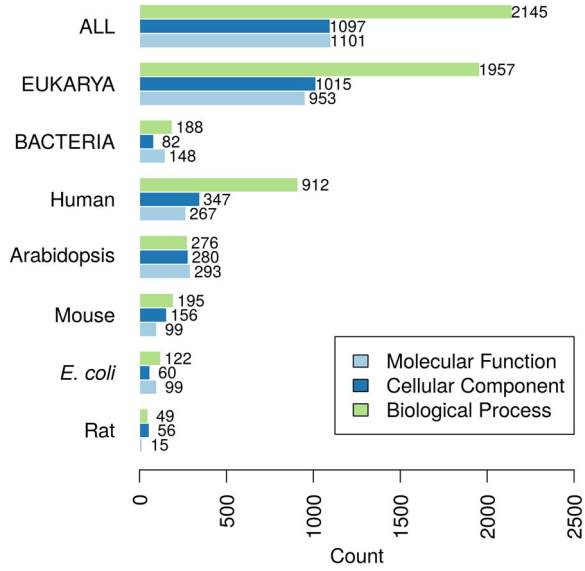


Fig. 6 Number of proteins in each benchmark species and ontology

CAFA1 top methods are more diverse than CAFA2 and CAFA3. The off-diagonal blocks shows that CAFA2 and CAFA3 methods are more similar with each other than with CAFA3 methods. It is clear that some methods are heavily based on the Naïve and BLAST baseline methods.

Participating teams also provided keywords that describe their approach to function prediction with their submissions. A list of keywords was given to the participants, listed in Additional file 1. Figure 8 shows the frequency of each keyword. In addition, we have weighted the frequency of the keywords with the prediction accuracy of the specific method. Machine learning and sequence alignment remain the most used approach by scientists predicting in all three ontologies. By raw count, machine learning is more popular than sequence in all three ontologies, but once adjusted by performance, their difference shrinks. In MFO, sequence alignment even overtakes machine learning as the most popular keyword after adjusting for performance. This indicates that methods that use sequence alignments are more helpful in predicting the correct function than the popularity of their use suggests.

Evaluation via molecular screening

Databases with proteins annotated by biocuration, such as UniProt knowledge base and UniProt Gene Ontology Annotation (GOA) database, have been the primary source of benchmarks in the CAFA challenges. New to CAFA3, we also evaluated the extent to which methods participating in CAFA could predict the results of genetic screens in model organisms done specifically for this project. Predicting GO terms for a protein (protein-centric) and predicting which proteins are associated with a given function (term-centric) are related but different computational problems: the former is a multi-label classification problem with a structured output, while the latter is a binary classification task. Predicting the results of a genome-wide screen for a single or a small number of functions fits the term-centric formulation. To see how well all participating CAFA methods perform term-centric predictions, we mapped the results from the protein-centric CAFA3 methods onto these terms. In addition, we held a separate CAFA challenge, CAFA-π, whose purpose was to attract additional submissions from algorithms that specialize in term-centric tasks.

We performed screens for three functions in three species, which we then used to assess protein function prediction. In the bacterium *Pseudomonas aeruginosa* and the fungus *Candida albicans*, we performed genome-wide screens capable of uncovering genes with two functions, biofilm formation (GO:0042710) and motility (for *P. aeruginosa* only) (GO:0001539), as described in the "Methods" section. In *Drosophila melanogaster*, we performed targeted assays, guided by previous CAFA submissions, of a selected set of genes and assessed whether or not they affected long-term memory (GO:0007616).

We discuss the prediction results for each function below in detail. The performance, as assessed by the genome-wide screens, was generally lower than in the protein-centric evaluations that were curation driven. We hypothesize that it may simply be more difficult to perform term-centric prediction for broad activities such as biofilm formation and motility. For *P. aeruginosa*, an existing compendium of gene expression data was already available [33]. We used the Pearson correlation over this collection of data to provide a complementary baseline to the standard BLAST approach used throughout CAFA. We found that an expression-based method outperformed the CAFA participants, suggesting that success on certain term-centric challenges will require the use of different types of data. On the other hand, the performance of the methods in predicting long-term memory in the Drosophila genome was relatively accurate.

Biofilm formation

In March 2018, there were 3019 annotations to biofilm formation (GO:0042710) and its descendent terms across all species, of which 325 used experimental evidence codes. These experimentally annotated proteins included 131 from the Candida Genome Database [34] for *C. albicans* and 29 for *P. aeruginosa*, the 2 organisms that we screened.

Of the 2746 genes we screened in the *Candida albicans* colony biofilm assay, 245 were required for the formation of wrinkled colony biofilm formation (Table 1). Of these, only 5 were already annotated in UniProt: *MOB*,

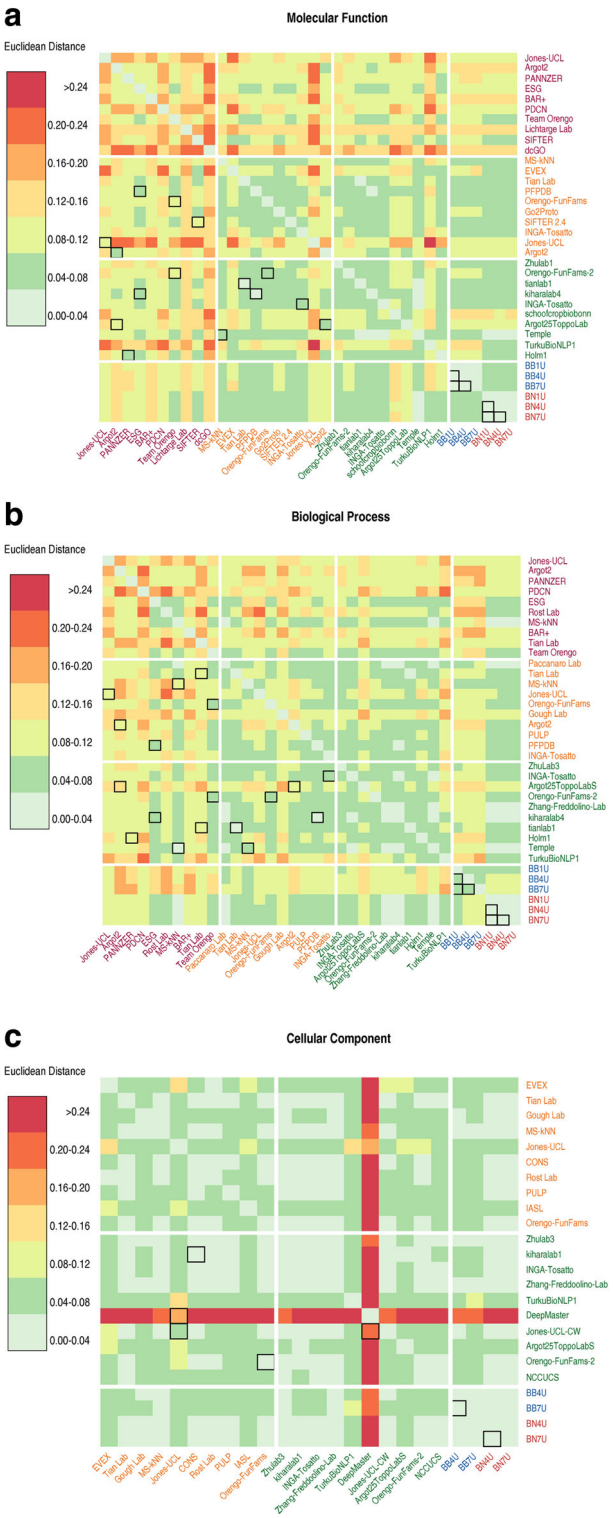


Fig. 7 Heatmap of similarity for the top 10 methods in CAFA1, CAFA2, and CAFA3. Similarity is represented by Euclidean distance of the prediction scores from each pair of methods, using the intersection set of benchmarks in the "Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2" section. The higher (darker red color) the euclidean distance, the less similar the methods are. Top 10 methods from each of the CAFA challenges are displayed and ranked by their performance in F_{max}. Cells highlighted by black borders are between a pair of methods that come from the same PI. **a**: Molecular Function; **b**: Biological Process; **c**: Cellular Component

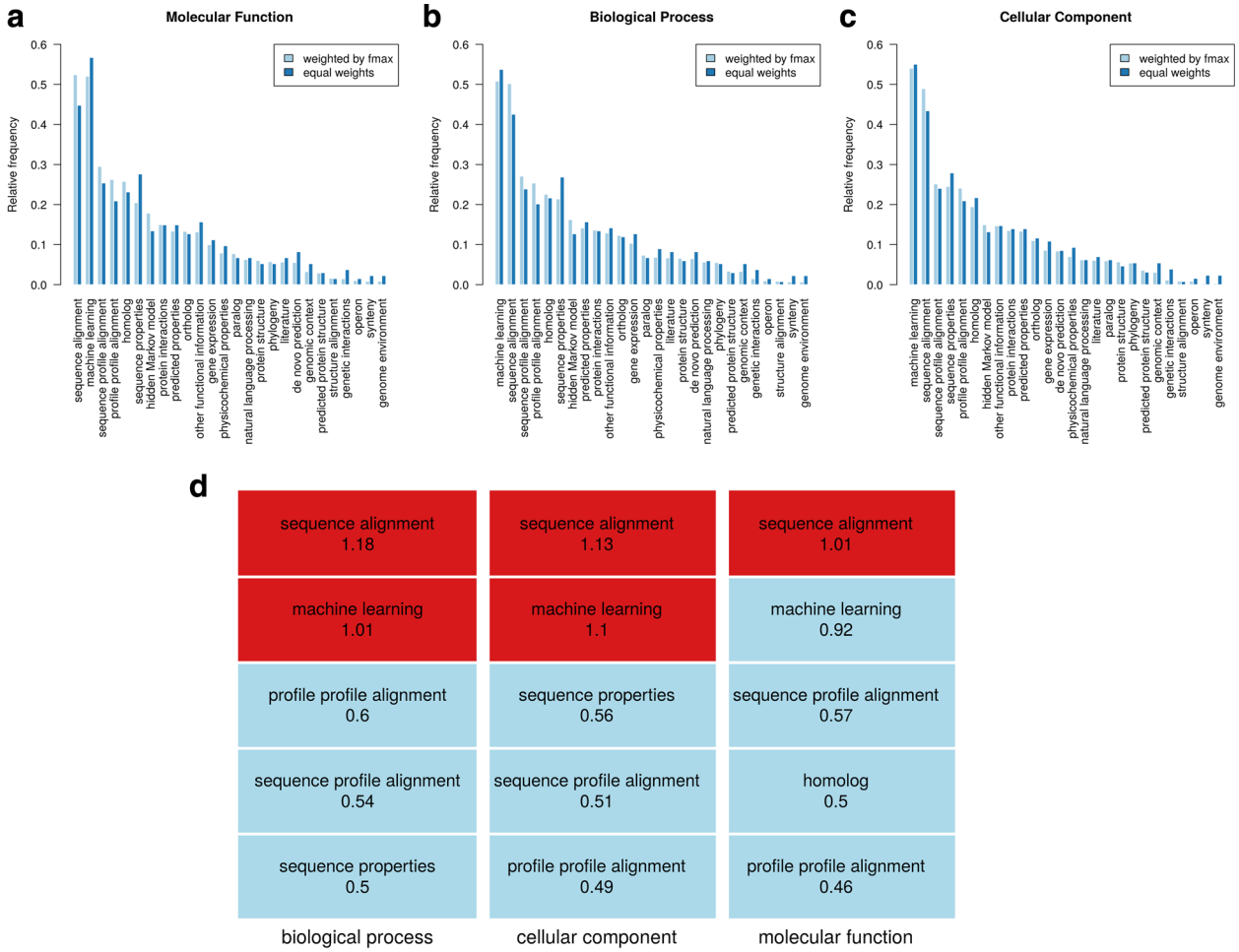


Fig. 8 Keyword analysis of all CAFA3 participating methods. **a**–**c**: both relative frequency of the keywords and weighted frequencies are provided for three respective GO ontologies. The weighted frequencies accounts for the performance of the the particular model using the given keyword. If that model performs well (with high F_{max}), then it gives more weight to the calculation of the total weighted average of that keyword. **d** shows the ratio of relative frequency between the F_{max}-weighted and equal-weighted. Red indicates the ratio is greater than one while blue indicates the ratio is less than one. Only the top five keywords ranked by ratio are shown. The larger the ratio, the more difference there is between the F_{max}-weighted and the equal-weighted

EED1 (*DEF1*), and *YAK1*, which encode proteins involved in hyphal growth, an important trait for biofilm formation [35–38]. Also, *NUP85*, a nuclear pore protein involved in early phase arrest of biofilm formation [39] and *VPS1*, contributes to protease secretion, filamentation, and biofilm formation [40]. Of the 2063 proteins that we did not find to be associated with biofilm formation, 29 were annotated with the term in the GOA database in *C. albicans*. Some of the proteins in this category highlight the need for additional information to GO term annotation. For example, Wor1 and the pheromone receptor are key for biofilm formation in strains under conditions in which the mating pheromone is produced [41], but not required in the monocultures of the commonly studied a/α mating type strain used here.

Table 1 Number of proteins in *Candida albicans* and *Pseudomonas aeruginosa* associated with the GO term "Biofilm formation" (GO:0042710) in the GOA databases versus experimental results

			GOA annotations	
C. albicans	Total, 2308		Unannotated	Annotated
	CAFA experiments	False	2034	29
		True	240	5
P. aeruginosa	Total, 4056		Unannotated	Annotated
	CAFA experiments	False	3491	25
		True	532	9

We used receiver operating characteristic (ROC) curves to measure the prediction accuracy. Area under ROC curves (AUROC) was used to compare the performance. AUROC is a common accuracy measure for classification problems where it evaluates how good a model is at distinguishing between the positive and negative classes. No method in CAFA-π or CAFA3 (not shown) exceeded an AUC of 0.60 on this term-centric challenge (Fig. 9) for either species. Performance for the best methods slightly exceeded a BLAST-based baselines. In the past, we have found that predicting BPO terms, such as biofilm formation, resulted in poorer method performance than predicting MFO terms. Many CAFA methods use sequence alignment as their primary source of information (the "Diversity of methods" section). For *Pseudomonas aeruginosa*, a pre-built expression compendium was available from prior work [33]. Where the compendium was available, simple gene expression-based baselines were the best-performing approaches. This suggests that successful term-centric prediction of biological processes may need to rely more heavily on information that is not sequence-based and, as previously reported, may require methods that use broad collections of gene expression data [42, 43].

Motility

In March 2018, there were 302,121 annotations for proteins with the GO term: cilium or flagellum-dependent cell motility (GO:0001539) and its descendent terms, which included cell motility in all eukaryotic (GO:0060285), bacterial (GO:0071973), and archaeal (GO:0097590) organisms. Of these, 187 had experimental

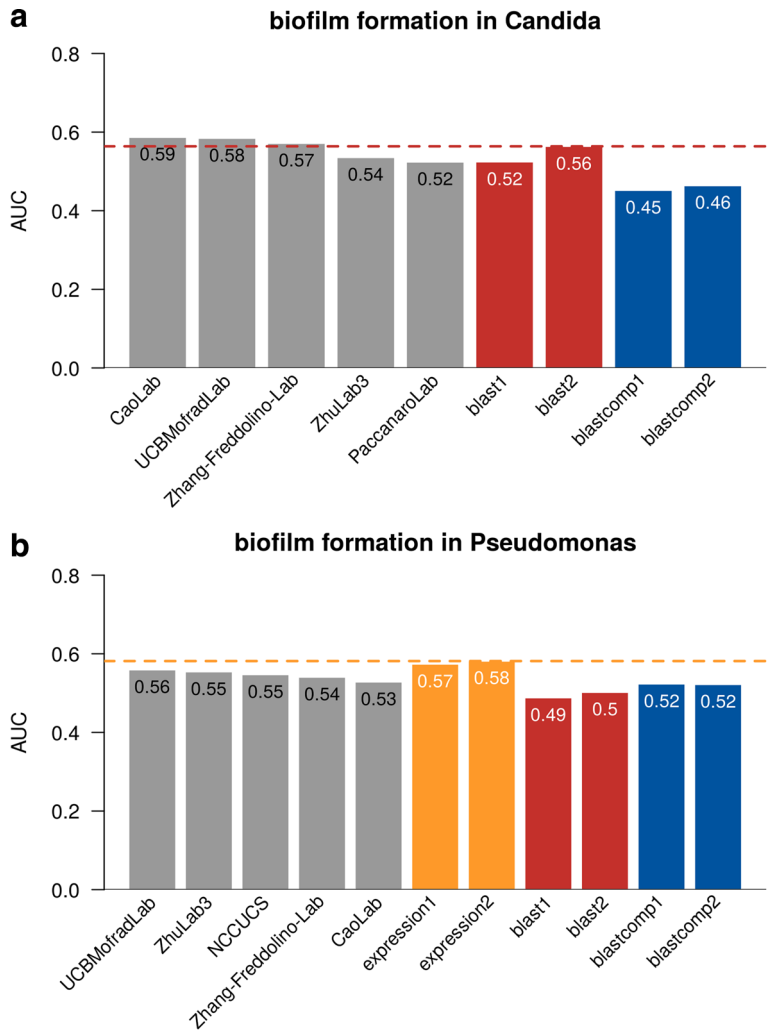


Fig. 9 AUROC of the top five teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. Four baseline models all based on BLAST were computed for *Candida*, while six baseline models were computed for *Pseudomonas*, including two based on expression profiles. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue, and expression are in yellow, see Table 3 for the description of the baselines

evidence codes, and the most common organism to have annotations was *P. aeruginosa*, on which our screen was performed (Additional file 1: Table S2).

As expected, mutants defective in the flagellum or its motor were defective in motility (*fliC* and other *fli* and *flg* genes). For some of the genes that were expected, but not detected, the annotation was based on the experiments performed in a medium different from what was used in these assays. For example, PhoB regulates motility but only when phosphate concentration is low [44]. Among the genes that were scored as defective in motility, some are known to have decreased motility due to over production of carbohydrate matrix material (*bifA*) [45], or the absence of directional swimming due to absence of chemotaxis functions (e.g., *cheW*, *cheA*) and others likely showed this phenotype because of a medium-specific requirement such as biotin (*bioA*, *bioC*, and *bioD*) [46]. Table 2 shows the contingency table for the number of proteins that are detected by our experiment versus GOA annotations.

The results from this evaluation were consistent with what we observed for biofilm formation. Many of the genes annotated as being involved in biofilm formation were identified in the screen. Others that were annotated as being involved in biofilm formation did not show up in the screen because the strain background used here, strain PA14, uses the exopolysaccharide matrix carbohydrate Pel [47] in contrast to the Psl carbohydrate used by another well-characterized strain, strain PAO1 [48, 49]. The *psl* genes were known to be dispensable for biofilm formation in the strain PA14 background, and this nuance highlights the need for more information to be taken into account when making predictions.

The CAFA-π methods outperformed our BLAST-based baselines but failed to outperform the expression-based baselines. Transferred methods from CAFA3 also did not outperform these baselines. It is important to note this consistency across terms, reinforcing the finding that term-centric prediction of biological processes is likely to require non-sequence information to be included.

Long-term memory in D. melanogaster

Prior to our experiments, there were 1901 annotations made in the long-term memory, including 283 experimental annotations. *Drosophila melanogaster* had the most annotated proteins of long-term memory with 217, while human has 7, as shown in Additional file 1: Table S3.

We performed RNAi experiments in *Drosophila melanogaster* to assess whether 29 target genes were associated with long-term memory (GO:0007616). Briefly, flies were exposed to wasps, which triggers a behavior that causes females to lay fewer eggs. The acute response is measured until 24 h post-exposure, and the long-term response is measured at 24 to 48 h post-exposure. RNAi was used to interfere with the expression of the 29 target genes in the mushroom body, a region of the fly brain associated with memory. Using this assay, we identified 3 genes involved in the perception of wasp exposure and 12 genes involved in the long-term memory. For details on the target selection and fly assay, see [29]. None of the 29 genes had an existing annotation in the GOA database. Because no genome-wide screen results were available, we did not release this as part of the CAFA-π and instead relied only on the transfer of methods that predicted the "long-term memory" at least once in *D. melanogaster* from CAFA3. Results from this assessment were more promising than our findings from the genome-wide screens in microbes (Fig. 10). Certain methods performed well, substantially exceeding the baselines.

Table 2 Number of proteins in *Pseudomonas aeruginosa* associated with function motility (GO:0001539) in the GOA databases versus experimental results

		GOA annotations	
Total, 3630		Unannotated	Annotated
CAFA experiments	False	3195	12
	True	403	21

Participation growth

The CAFA challenge has seen growth in participation, as shown in Fig. 11. To cope with the increasingly large data size, CAFA3 utilized the Synapse [50] online platform for submission. Synapse allowed for easier access for participants, as well as easier data collection for the organizers. The results were also released to the individual teams via this online platform. During the submission process, the online platform also allows for customized format checkers to ensure the quality of the submission.

Methods

Benchmark collection

In CAFA3, we adopted the same benchmark generation methods as CAFA1 and CAFA2, with a similar timeline (Fig. 12). The crux of a time-delayed challenge is the annotation growth period between time t_0 and t_1. All target proteins that have gained experimental annotation during this period are taken as benchmarks in all three ontologies. "No knowledge" (NK, no prior experimental annotations) and "Limited knowledge" (LK, partial prior experimental annotations) benchmarks were also distinguished based on whether the newly gained experimental annotation is in an ontology that already have experimental annotations or not. Evaluation results in Figs. 3 and 4 are made using the No knowledge benchmarks. Evaluation results on the Limited knowledge benchmarks are

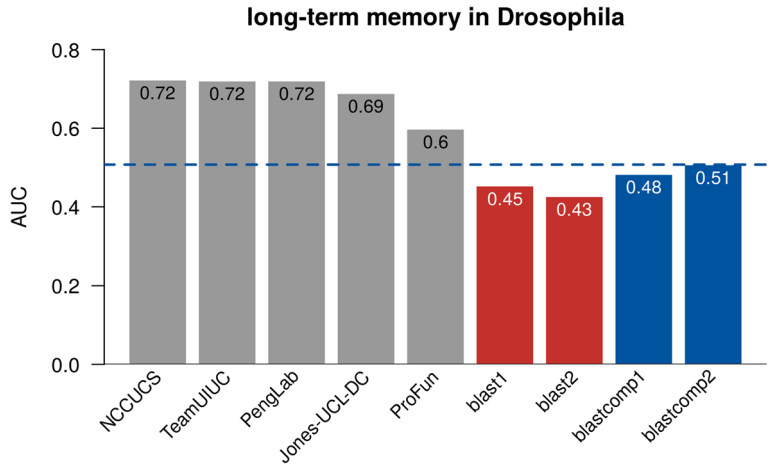


Fig. 10 AUROC of top five teams in CAFA3. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red and BLAST computational methods are in blue, see Table 3 for the description of the baselines

shown in Additional file 1: Figure S3. For more information regarding NK and LK designations, please refer to the Additional file 1 and the CAFA2 paper [26].

After collecting these benchmarks, we performed two major deletions from the benchmark data. Upon inspecting the taxonomic distribution of the benchmarks, we noticed a large number of new experimental annotations from *Candida albicans*. After consulting with UniProt-GOA, we determined these annotations have already existed in the Candida Genome Database long before 2018 but were only recently migrated to GOA. Since these annotations were already in the public domain before the CAFA3 submission deadline, we have deleted any annotation from *Candida albicans* with an assigned date prior to our CAFA3 submission deadline. Another major change is the deletion of any proteins with only a protein-binding (GO:0005515) annotation. Protein binding is a highly generalized function description, does not provide more specific information about the actual function of a protein, and in many cases may indicate a non-functional, non-specific binding. If it is the only annotation that a protein has gained, then it is hardly an advance in our understanding of that protein; therefore, we deleted these annotations from our benchmark set. Annotations with a depth of 3 make up almost half of all annotations in MFO before the removal (Additional file 1: Figure S15B). After the removal, the most frequent annotations became of depth 5 (Additional file 1: Figure S15A). In BPO, the most frequent annotations are of depth 5 or more, indicating a healthy increase of specific GO terms being added to our annotation database. In CCO, however, most new annotations in our benchmark set are of depths 3, 4, and 5 (Additional file 1: Figure S15). This difference could partially explain why the same computational methods perform very differently in different ontologies and benchmark sets. We have also calculated the total

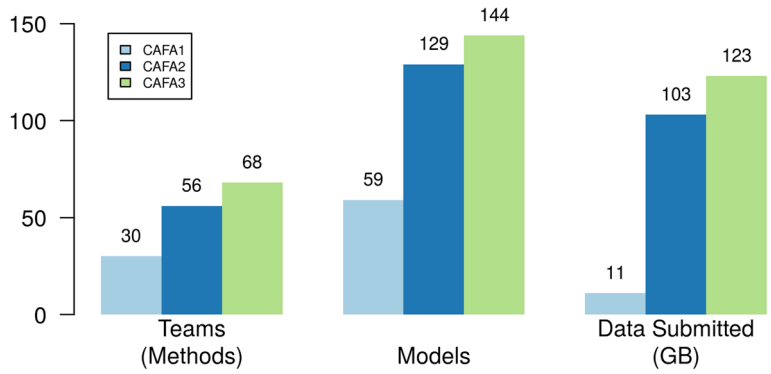


Fig. 11 CAFA participation has been growing. Each principal investigator is allowed to head multiple teams, but each member can only belong to one team. Each team can submit up to three models

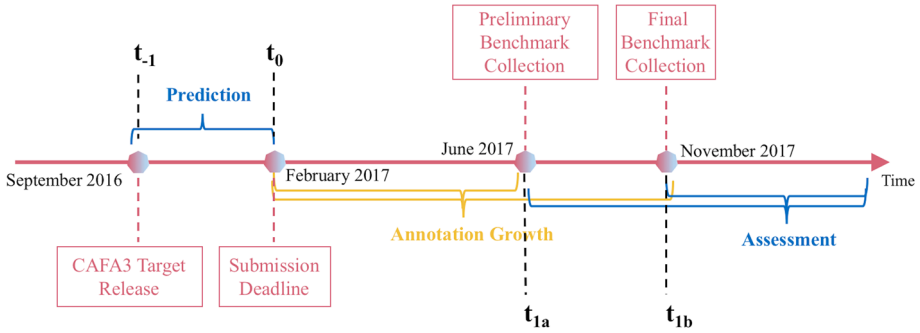


Fig. 12 CAFA3 timeline

information content per protein for the benchmark sets shown in Additional file 1: Figure S16. Taxonomic distributions of the proteins in our final benchmark set are shown in Fig. 6.

Additional analyses were performed to assess the characteristics of the benchmark set, including the overall information content of the terms being annotated.

Protein-centric evaluation

Two main evaluation metrics were used in CAFA3, the F_{max} and the S_{min}. The F_{max} based on the precision-recall curve (Fig. 3), while the S_{min} is based on the remaining uncertainty/missing information (RU-MI) curve as described in [51] (Fig. 4), where S stands for semantic distance. The shortest semantic distance across all thresholds is used as the S_{min} metric. The RU-MI curve takes into account the information content of each GO term in addition to counting the number of true positives, false positives, etc., see Additional file 1 for the precise definition of F_{max} and S_{min}. The information theory-based evaluation metrics counter the high-throughput low-information annotations such as protein binding, but down-weighing these terms according to their information content, as the ability to predict such non-specific functions are not as desirable and useful and the ability to predict more specific functions.

The two assessment modes from CAFA2 were also used in CAFA3. In the partial mode, predictions were evaluated only on those benchmarks for which a model made at least one prediction. The full evaluation mode evaluates all benchmark proteins, and methods were penalized for not making predictions. Evaluation results in Figs. 3 and 4 are made using the full evaluation mode. Evaluation results using the partial mode are shown in Additional file 1: Figure S2.

Two baseline models were also computed for these evaluations. The Naïve method assigns the term frequency as the prediction score for any protein, regardless of any protein-specific properties. BLAST was based on the results using the Basic Local Alignment Search Tool (BLAST) software against the training database [52]. A term will be predicted as the highest local alignment sequence identity among all BLAST hits annotated from the training database. Both of these methods were trained on the experimentally annotated proteins and their sequences in Swiss-Prot [53] at time t_0.

Microbe screens

To assess the matrix production, we used mutants from the PA14 NR collection [54]. Mutants were transferred from the − 80 °C freezer stock using a sterile 48-pin multiprong device into 200 μl LB in a 96-well plate. The cultures were incubated overnight at 37 °C, and their OD600 was measured to assess growth. Mutants were then transferred to tryptone agar with 15 g of tryptone and 15 g of agar in 1L amended with Congo red (Aldrich, 860956) and Coomassie brilliant blue (J.T. Baker Chemical Co., F789-3). Plates were incubated at 37 °C overnight followed by 4-day incubation at room temperature to allow the wrinkly phenotype to develop. Colonies were imaged and scored on day 5. To assess motility, mutants were revived from freezer stocks as described above. After overnight growth, a sterile 48-pin multiprong transfer device with a pin diameter of 1.58 mm was used to stamp the mutants from the overnight plates into the center of swim agar made with M63 medium with 0.2% glucose and casamino acids and 0.3% agar). Care was taken to avoid touching the bottom of the plate. Swim plates were incubated at room temperature (19–22 °C) for approximately 17 h before imaging and scoring. Experimental procedures in *P. aeruginosa* to determine proteins that are associated with the two functions in CAFA-π are shown in Fig. 13.

Biofilm formation in *Candida albicans* was assessed in single gene mutants from the Noble [55] and GRACE [56] collections. In the Noble Collection, mutants of *C. albicans* have had both copies of the candidate gene deleted. Most of the mutants were created in biological duplicate. From this collection, 1274 strains corresponding to 653 unique genes were screened. The GRACE Collection

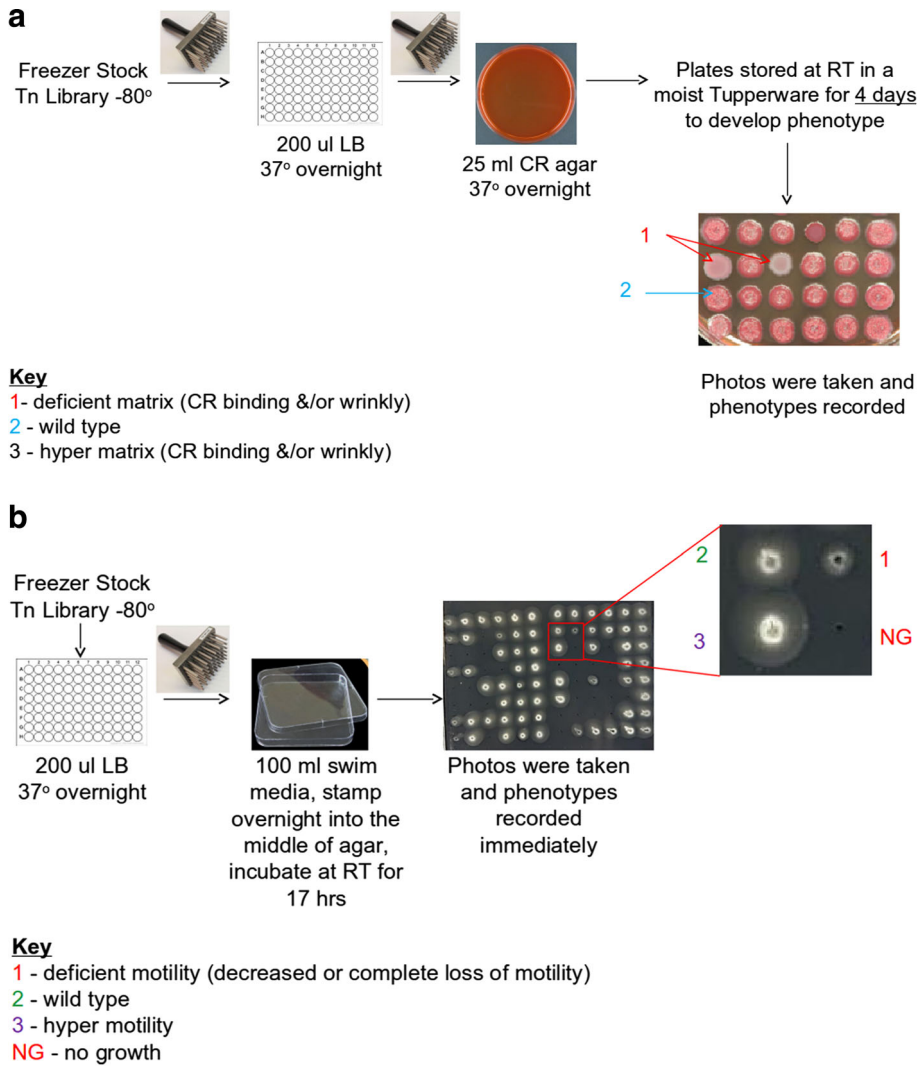


Fig. 13 Experimental procedure of determining genes associated with the functions biofilm formation (**a**) and motility (**b**) in *P. aeruginosa*

provided mutants with one copy of each gene deleted and the other copy placed under the control of a doxycycline-repressible promoter. To assay these strains, we used a medium supplemented with 100 μg/ml doxycycline strains, when rendered them functional null mutants. We screened 2348 mutants from the GRACE Collection, 255 of which overlapped with mutants in the Noble Collection, for 2746 total unique mutants screened in total. To assess the defects in biofilm formation or biofilm-related traits, we performed 2 assays: (1) colony morphology on agar medium and (2) biofilm formation on a plastic surface (Fig. 14). For both of these assays, we used Spider medium, which was designed to induce hyphal growth in *C. albicans* [57] and which promotes biofilm formation [39]. Strains were first replicated from frozen 96-well plates to YPD agar plates. Strains were then replicated from YPD agar to YPD broth and grown overnight at 30 °C. From YPD broth, strains were introduced onto Spider agar plates and into 96-well plates of Spider broth. When strains from the GRACE Collection were assayed, 100 μg/ml doxycycline was included in the agar and broth, and aluminum foil was used to protect the media from light. Spider agar plates inoculated with *C. albicans* mutants were incubated at 37 °C for 2 days before colony morphologies were scored. Strains in Spider broth were shaken at 225 rpm at 37 °C for 3 days and then assayed for biofilm formation at the air-liquid interface as follows. First, broth was removed by slowly tilting the plates and pulling the liquid away by running a gloved hand over the surface. Biofilms were stained by adding 100 μl of 0.1 percent crystal violet dye in water to each well of the plate. After 15 min, plates were gently washed in three

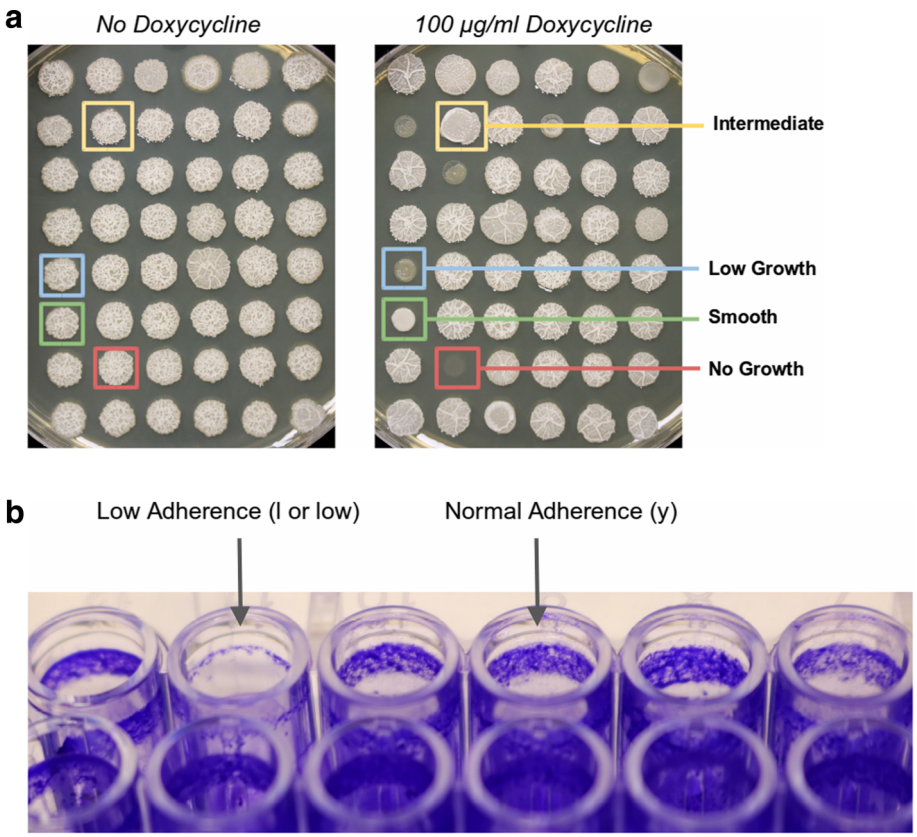


Fig. 14 a: different phenotypes in response to doxycycline treatment: low growth, smooth, no growth and intermediate. **b**: adherence phenotypes. See text for details

baths of water to remove dye without disturbing biofilms. To score biofilm formation for agar plates, colonies were scored by eye as either smooth, intermediate, or wrinkled. A wild-type colony would score wrinkled, and mutants with intermediate or smooth appearance were considered defective in colony biofilm formation. For biofilm formation on a plastic surface, the presence of a ring of cell material in the well indicated normal biofilm formation, while low or no ring formation mutants were considered defective. Genes whose mutations resulted defects in both or either assay were considered true for biofilm function. A complete list of the mutants identified in the screens is available in Additional file 1: Table S1.

A protein is considered true in the biofilm function, if its mutant phenotype is smooth or intermediate under doxycycline.

Term-centric evaluation

The evaluations of the CAFA-π methods were based on the experimental results in the "Microbe screens" section. We adopted $F_{\max}$ based on both precision-recall curves and area under ROC curves. There are a total of six baseline methods, as described in Table 3.

Discussion

Since 2010, the CAFA community has been the home to a growing group of scientists across the globe sharing the goal of improving computational function prediction. CAFA has been advancing this goal in three ways. First, through independent evaluation of computational methods against the set of benchmark proteins, thus providing a direct comparison of the methods' reliability and performance at a given time point. Second, the challenge assesses the quality of the current state of the annotations, whether they are made computationally or not, and is set up to reliably track it over time. Finally, as described in this work, CAFA has started to drive the creation of new experimental annotations by facilitating synergies between different groups of researchers interested in function of biological macromolecules. These annotations not only represent new biological discoveries, but simultaneously serve to provide benchmark data for rigorous method evaluation.

CAFA3 and CAFA-π feature the latest advances in the CAFA series to create advanced and accurate methods for protein function prediction. We use the repeated nature of the CAFA project to identify certain trends

Table 3 Baseline methods in term-centric evaluation of protein function prediction

	Model number	Training data	Score assignment
Expression	1	Gene expression compendium for *P. aeruginosa PAO1*	Highest correlation score out of all pairwise correlations
	2		Top 10 average correlation score
BLAST	1	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	Highest sequence identity out of all pairwise BLASTp hits
blastcomp	1	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	

via historical assessments. The analysis revealed that the performance of CAFA methods improved dramatically between CAFA1 and CAFA2. However, the protein-centric results for CAFA3 are mixed when compared to historical methods. Though the best-performing CAFA3 method outperformed the top CAFA2 methods (Fig. 1), this was not consistently true for other rankings. Among all 3 CAFA challenges, CAFA2 and CAFA3 methods inhabit the top 12 places in MFO and BPO. Between CAFA2 and CAFA3, the performance increase is more subtle. Based on the annotations of methods (Additional file 1), many of the top-ranking methods are improved versions of the methods that have been evaluated in CAFA2. Interestingly, the top-performing CAFA3 method, which consistently outperformed the methods from all past CAFAs in the major categories (GOLabeler [23]), utilized 5 component classifiers trained from different features; those included GO term frequency, sequence alignment, amino acid trigram, domains, motifs, and biophysical properties. It performs best in the Molecular Function Ontology, where sequence features perform best. Another method which did not participate in CAFA3 yet seems to perform well under CAFA parameters is NetGO [58], which utilizes the information from STRING, a network association database [59] in addition to sequence information. Taken together, the strong predictive performance of mRNA co-expression data (Figs. 9 and 15) leads us to hypothesize that including more varied sources of data can lead to additional large improvements in protein function prediction. We are looking forward to testing this hypothesis in future CAFA challenges. It should be noted that CAFA uses both $F_{\max}$ and $S_{\min}$.

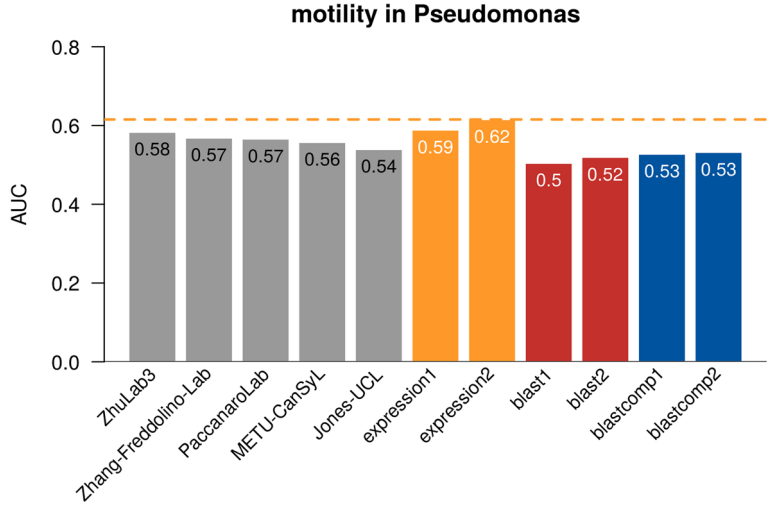


Fig. 15 AUROC of top 5 teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue and expression are in yellow. See Table 3 for description of the baselines

F_{max}'s strength lies in its interpretability, as it is simply the maximum F_1 given for each model. At the same time, precision/recall-based assessment does not capture the hierarchical nature of ontologies or the differences in information content between different GO terms. For that reason, we also use the S_{min} score which incorporates information content, but is somewhat less interpretable than F_{max} and less robust to the problems of incomplete annotation [60, 61]. Additionally, since the information content of a GO term is derived from its frequency in the corpus [62], it is somewhat malleable depending on the corpus from which it is derived. We therefore use both measures for scoring, to achieve a more comprehensive picture of the models' performance.

For this iteration of CAFA, we performed genome-wide screens of phenotypes in *P. aeruginosa* and *C. albicans* as well as a targeted screen in *D. melanogaster*. This not only allowed us to assess the accuracy with which methods predict genes associated with select biological processes, but also to use CAFA as an additional driver for new biological discovery. Note that high-throughput screening for a single phenotype should be interpreted with caution as the phenotypic effect may be the result of pleiotropy, and the phenotype in question may be expressed as part of a set of other phenotypes. The results of genome-wide screenings typically lack context for the observed phenotypic effects, and each genotype-phenotype association should be examined individually to ascertain how immediate is the phenotypic effect from the seeming genotypic cause.

In sum, our experimental work identified more than a thousand new functional annotations in three highly divergent species. Though all screens have certain limitations, the genome-wide screens also bypass questions of biases in curation. This evaluation provides key insights: CAFA3 methods did not generalize well to selected terms. Because of that, we ran a second effort, CAFA-π, in which participants focused solely on predicting the results of these targeted assays. This targeted effort led to improved performance, suggesting that when the goal is to identify genes associated with a specific phenotype, tuning methods may be required.

For CAFA evaluations, we have included both Naïve and sequence-based (BLAST) baseline methods. For the evaluation of *P. aeruginosa* screen results, we were also able to include a gene expression baseline from a previously published compendium [33]. Intriguingly, the expression-based predictions outperformed the existing methods for this task. In future CAFA efforts, we will include this type of baseline expression-based method across evaluations to continue to assess the extent to which this data modality informs gene function prediction. The results from the CAFA3 effort suggest that gene expression may be particularly important for successfully predicting term-centric biological process annotations.

The primary takeaways from CAFA3 are as follows: (1) genome-wide screens complement annotation-based efforts to provide a richer picture of protein function prediction; (2) the best-performing method was a new method, instead of a light retooling of an existing approach; (3) gene expression, and more broadly, systems data may provide key information to unlocking biological process predictions, and (4) performance of the best methods has continued to improve. The results of the screens released as part of CAFA3 can lead to a re-examination of approaches which we hope will lead to improved performance in CAFA4.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1835-8.

Additional file 1: Additional figures and tables referenced in the article.
Additional file 2: Review History.

Acknowledgements

NZ and IF acknowledge the invaluable input from Michael C Gerten and Shatabdi Sen and all members of the Friedberg Lab for the ongoing support for stimulating discussions.

Review history

The review history is available as Additional file 2.

Authors' contributions

The experiment was designed by IF, PR, CSG, SDM, COD, MJM, and NZ. NZ, YJ, MNH, HNN, AJL, and LD performed the computational analyses. NZ, SDM, and TM were responsible for managing the participants' submissions to the CAFA challenge. KAL, AWC, and DAH performed the experimental work in *C. albicans* and *P. aeruginosa*. BZK and GB performed the experimental work in *D. melanogaster*. CJJ, MJM, COD, and GG provided the novel biocurated data for the benchmarks and incorporated data into UniprotKB. All other co-authors developed the computational function prediction methods participating in the challenge, performed the computational protein function predictions, and submitted the results for analysis in CAFA3. NZ, IF, CSG, DAH, and PR wrote the manuscript. All authors have read and approved the final manuscript.

Funding

The work of IF was funded, in part, by the National Science Foundation award DBI-1458359. The work of CSG and AJL was funded, in part, by the National Science Foundation award DBI-1458390 and GBMF 4552 from the Gordon and Betty Moore Foundation. The work of DAH and KAL was funded, in part, by the National Science Foundation award DBI-1458390, National Institutes of Health NIGMS P20 GM113132, and the Cystic Fibrosis Foundation CFRDP STANTO19R0. The work of AP, HY, AR, and MT was funded by BBSRC grants BB/K004131/1, BB/F00964X/1 and BB/M025047/1, Consejo Nacional de Ciencia y Tecnología Paraguay (CONACyT) grants 14-INV-088 and PINV15-315, and NSF Advances in BioInformatics grant 1660648. The work of JC was partially supported by an NIH grant (R01GM093123) and two NSF grants (DBI 1759934 and IIS1763246). ACM acknowledges the support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2155 "RESIST" - Project ID 39087428. DK acknowledges the support from the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941). PB acknowledges the support from the National Institutes of Health (R01GM60595). GB and BZK acknowledge the support from the National Science Foundation (NSF 1458390) and NIH DP1MH110234. FS was funded by the ERC StG 757700 "HYPER-INSIGHT" and by the Spanish Ministry of Science, Innovation and Universities grant BFU2017-89833-P. FS further acknowledges the funding from the Severo Ochoa award to the IRB Barcelona. TS was funded by the Centre of Excellence project "BioProspecting of Adriatic

Sea", co-financed by the Croatian Government and the European Regional Development Fund (KK.01.1.1.01.0002). The work of SK was funded by ATT Tieto käyttöön grant and Academy of Finland. JB and HM acknowledge the support of the University of Turku, the Academy of Finland and CSC – IT Center for Science Ltd. TB and SM were funded by the NIH awards UL1 TR002319 and U24 TR002306. The work of CZ and ZW was funded by the National Institutes of Health R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of PWR was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. PR acknowledges NSF grant DBI-1458477. PT acknowledges the support from Helsinki Institute for Life Sciences. The work of AJM was funded by the Academy of Finland (No. 292589). The work of FZ and WT was funded by the National Natural Science Foundation of China (31671367, 31471245, 91631301) and the National Key Research and Development Program of China (2016YFC1000505, 2017YFC0908402]. CS acknowledges the support by the Italian Ministry of Education, University and Research (MIUR) PRIN 2017 project 2017483NH8. SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). PLF and RLH were supported by the National Institutes of Health NIH R35-GM128637 and R00-GM097033. JG, DTJ, CW, DC, and RF were supported by the UK Biotechnology and Biological Sciences Research Council (BB/N019431/1, BB/L020505/1, and BB/L002817/1) and Elsevier. The work of YZ and CZ was funded in part by the National Institutes of Health award GM083107, GM116960, and AI134678; the National Science Foundation award DBI1564756; and the Extreme Science and Engineering Discovery Environment (XSEDE) award MCB160101 and MCB160124. The work of BG, VP, RD, NS, and NV was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 173001. The work of YWL, WHL, and JMC was funded by the Taiwan Ministry of Science and Technology (106-2221-E-004-011-MY2). YWL, WHL, and JMC further acknowledge the support from "the Human Project from Mind, Brain and Learning" of the NCCU Higher Education Sprout Project by the Taiwan Ministry of Education and the National Center for High-performance Computing for computer time and facilities. The work of IK and AB was funded by Montana State University and NSF Advances in Biological Informatics program through grant number 0965768. BR, TG, and JR are supported by the Bavarian Ministry for Education through funding to the TUM. The work of RB, VG, MB, and DCEK was supported by the Simons Foundation, NIH NINDS grant number 1R21NS103831-01 and NSF award number DMR-1420073. CJJ acknowledges the funding from a University of Illinois at Chicago (UIC) Cancer Center award, a UIC College of Liberal Arts and Sciences Faculty Award, and a UIC International Development Award. The work of ML was funded by Yad Hanadiv (grant number 9660 /2019). The work of OL and IN was funded by the National Institute of General Medical Science of the National Institute of Health through GM066099 and GM079656. Research Supporting Plan (PSR) of University of Milan number PSR2018-DIP-010-MFRAS. AWV acknowledges the funding from the BBSRC (CASE studentship BB/M015009/1). CD acknowledges the support from the Swiss National Science Foundation (150654). CO and MJM are supported by the EMBL-European Bioinformatics Institute core funds and the CAFA BBSRC BB/N004876/1. GG is supported by CAFA BBSRC BB/N004876/1. SCET acknowledges funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778247 (IDPfun) and from COST Action BM1405 (NGP-net). SEB was supported by NIH/NIGMS grant R01 GM071749. The work of MLT, JMR, and JMF was supported by the National Human Genome Research Institute of the National of Health, grant numbers U41 HG007234. The work of JMF and JMR was also supported by INB Grant (PT17/0009/0001 - ISCIII-SGEFI / ERDF). VA acknowledges the funding from TUBITAK EEEAG-116E930. RCA acknowledges the funding from KanSil 2016K121540. GV acknowledges the funding from Università degli Studi di Milano - Project "Discovering Patterns in Multi-Dimensional Data" and Project "Machine Learning and Big Data Analysis for Bioinformatics". SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). RY and SY are supported by the 111 Project (NO. B18015), the key project of Shanghai Science & Technology (No. 16JC1420402), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), and ZJLab. ST was supported by project Ribes Network POR-FESR 3S4H (No. TOPP-ALFREVE18-01) and PRID/SID of University of Padova (No. TOPP-SID19-01). CZ and ZW were supported by the NIGMS grant R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of MK and RH was supported by the funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01-01 and URF/1/3790-01-01. The work of SDM is funded, in part, by NSF award DBI-1458443.

Availability of data and materials

Data repository: A data repository providing all additional data, analyses, and all anonymous prediction results for all methods are available at https://figshare.com/articles/Supplementary_data/8135393/3 [63]. Code: The assessment software used in this paper is available under GNU-GPLv3 license on GitHub in both Matlab [64] (https://doi.org/10.5281/zenodo.3403452) and Python [65](http://doi.org/10.5281/zenodo.3401694).

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA. [2]Program in Bioinformatics and Computational Biology, Ames, IA, USA. [3]Indiana University Bloomington, Bloomington, Indiana, USA. [4]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA, USA. [5]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA. [6]Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [7]Department of Molecular and Systems Biology, Hanover, NH, USA. [8]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [9]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom. [10]Program in Computer Science, Ames, IA, USA. [11]Department of Computer Engineering, Hacettepe University, Ankara, Turkey. [12]Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey. [13]Department of Computer Engineering, Iskenderun Technical University, Hatay, Turkey. [14]CanSyL, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey. [15]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [16]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA. [17]Achira Labs, Bangalore, India. [18]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [20]INB Coordination Unit, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Catalonia, Spain. [21](former) INB GN2, Structural and Computational Biology Programme, Spanish National Cancer Research Centre, Barcelona, Catalonia, Spain. [22]Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia. [23]Molecular Cell Biomechanics Laboratory, Departments of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [24]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Berkeley, CA, USA. [25]Departments of Bioengineering and Mechanical Engineering, Berkeley, CA, USA. [26]Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy. [27]National Research Council, IBIOM, Bologna, Italy. [28]University of Bonn: INRES Crop Bioinformatics, Bonn, North Rhine-Westphalia, Germany. [29]INRES Crop Bioinformatics, University of Bonn, Bonn, Germany. [30]Gianforte School of Computing, Montana State University, Bozeman, Montana, USA. [31]University of Bristol, Computer Science, Bristol, Bristol, United Kingdom. [32]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Brunswick, Germany. [33]RESIST, DFG Cluster of Excellence 2155, Brunswick, Germany. [34]Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel, Brussels, Belgium. [35]Machine Learning Group, Université libre de Bruxelles, Brussels, Belgium. [36]Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium. [37]European Molecular Biolo gy Labora tory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [38]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom. [39]University of Kent, School of Computing, Canterbury, United Kingdom. [40]School of Biosciences, University of Kent, Canterbury, Kent, United Kingdom. [41]University of Missouri, Computer Science, Columbia, Missouri, USA. [42]Department of Electrical Engineering and Computer Science, University

of Missouri, Columbia, MO, USA. [43]University of Miami, Coral Gables, Florida, USA. [44]Centre for Systems and Synthetic Biology, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, United Kingdom. [45]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Galway, Ireland. [46]Technical University of Munich, Garching, Germany. [47]Faculty for Informatics, Garching, Germany. [48]Department for Bioinformatics and Computational Biology, Garching, Germany. [49]School of Computing Sciences and Computer Engineering, Hattiesburg, Mississippi, USA. [50]Institute of Biotechnology, Helsinki Institute of Life Sciences, University of Helsinki, Finland, Helsinki, Finland. [51]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [52]Compugen Ltd., Holon, Israel. [53]Baylor College of Medicine, Department of Biochemistry and Molecular Biology, Houston, TX, USA. [54]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, USA. [55]National TsingHua University, Hsinchu, Taiwan. [56]Department of Electrical Engineering in National Tsing Hua University, Hsinchu City, Taiwan. [57]The Hebrew University of Jerusalem, Jerusalem, Israel. [58]University of California San Diego, San Diego Supercomputer Center, La Jolla, California, USA. [59]Department of Computational Biology and Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [60]Department of Genetics, Evolution & Environment, and Department of Computer Science, University College London, London, UK. [61]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [62]Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia. [63]Jozef Stefan Institute, Ljubljana, Slovenia. [64]Jozef Stefan International Postgraduate School, Ljubljana, Slovenia. [65]Research Department of Structural and Molecular Biology, University College London, London, England. [66]Research Department of Structural and Molecular Biology, University College London, London, United Kingdom. [67]Oxford Brookes University, Department of Health and Life Sciences, London, UK. [68]University College London, Department of Computer Science, London, United Kingdom. [69]The Francis Crick Institute, Biomedical Data Science Laboratory, London, United Kingdom. [70]Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom. [71]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [72]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain. [73]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [74]Università degli Studi di Milano - Computer Science Department - AnacletoLab, Milan, Milan, Italy. [75]Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Universite de Montpellier, Montpellier, France. [76]Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [77]University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France. [78]University of Lorraine, Nancy, Lorraine, France. [79]Inria, Nancy, France. [80]Department of Biology, New York University, New York, NY, USA. [81]NYU Center for Data Science, New York, NY 10010, USA. [82]Flatiron Institute, CCB, 10010 New York, NY, USA. [83]Center for Computational Biology (CCB), Flatiron Institute, Simons Foundation, New York, New York, USA. [84]Center for Data Science, New York University, New York 10011, NY, USA. [85]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [86]Department of Molecular Medicine, University of Padova, Padova, Italy. [87]Department of Biology - University of Padova, Padova, Italy. [88]CNR Institute of Neuroscience, Padova, Italy. [89]Department of Biomedical Sciences, University of Padua, Padova, Italy. [90]Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar, Khyber Pakhtoonkhwa, Pakistan. [91]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [92]University of California, Riverside, Philadelphia, PA, USA. [93]Department of Biology, Brigham Young University, Provo, UT, USA. [94]Bioinformatics Research Group, Provo, UT, USA. [95]School of Biological Sciences, University of Reading, Reading, England, United Kingdom. [96]Department of Pharmaceutical Chemistry, San Francisco, CA, USA. [97]UC Berkeley - UCSF Graduate Program in Bioengineering, University of California, San Francisco 94158, CA, USA. [98]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco 94158, CA, USA. [99]Research and Innovation Center, Edmund Mach Foundation, 38010 San Michele all'Adige, Italy. [100]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, Shanghai, China. [101]Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [102]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China. [103]Institute of Science and Technology for Brain-Inspired Intelligence and Shanghai Institute of Artificial Intelligence Algorithms, Fudan University, Shanghai, China. [104]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China. [105]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [106]Department of Pediatrics, Brain Tumor Center, Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [107]Pacific Lutheran University, Department of Computer Science, Tacoma, WA, USA. [108]Department of Computer Science, National Chengchi University, Taipei, Taiwan. [109]Okinawa Institute of Science and Technology, Tancha, Okinawa, Japan. [110]Tel Aviv University, Tel Aviv, Israel. [111]Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Jeddah, Saudi Arabia. [112]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [113]Computer, Electrical and Mathematical Sciences Engineering Division (CEMSE), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [114]Politecnico di Torino, Control and Computer Engineering Department, Torino, TO, Italy. [115]University of Turku, Department of Future Technologies, Turku NLP Group, Turku, Finland. [116]University of Turku Graduate School (UTUGS), Turku, Finland. [117]University of Turku, Turku, Finland. [118]Turku Centre for Computer Science (TUCS), Turku, Finland. [119]Department of Future Technologies, Faculty of Science and Engineering, University of Turku, FI-20014 Turku, Finland. [120]Turku Centre for Computer Science (TUCS), Agora, Vesilinnantie 3, FI-20500, Turku, Finland. [121]University of Turku, Department of Future Technologies, Turku, Finland. [122]Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA. [123]Department of Pediatrics, University of Cincinnati, Cincinnati 45229, OH, USA. [124]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [125]Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia. [126]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [127]SIB Swiss Institute of Bioinformatics, Zurich, Switzerland. [128]Department of Computer Science, Colorado State University, Fort Collins, CO, USA. [129]Department of Informatics, Bioinformatics & Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [130]Institute for Food and Plant Sciences WZW, Technische Universität München, Freising, Germany. [131]University of California, Berkeley, CA, USA. [132]Biological Sciences, University of Illinois at Chicago, Chicago, Illinois, USA. [133]Geisel School of Medicine at Dartmouth, Department of Molecular and Systems Biology, Hanover, NH, USA. [134]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [135]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, Pennsylvania, USA. [136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA.

Received: 16 May 2019 Accepted: 24 September 2019
Published online: 19 November 2019

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
3. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol. 2013;9(5):1003063.
4. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. 2003;60(12):2637–50.
5. Friedberg I. Automated protein function prediction–the genomic challenge. Brief Bioinform. 2006;7(3):225–42.
6. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3:88.
7. Rentzsch R, Orengo CA. Protein function prediction–the power of multiplicity. Trends Biotechnol. 2009;27(4):210–9.
8. Shehu A, Barbara D, Molloy K. A survey of computational methods for protein function predictions. Cham: Springer; 2016, pp. 225–98.
9. Cozzetto D, Jones DT. Computational methods for annotation transfers from sequence. Methods Mol Biol. 2017;1446:55–67.

10. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA. 1999;96(8):4285–8.
11. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol. 2002;319(5):1257–65.
12. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. J Comput Biol. 2003;10(6):947–60.
13. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA. 2004;101(41):14754–9.
14. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005;21(21 Suppl 1):302–10.
15. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005;1(5):45.
16. Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005;6:247.
17. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 2006;15(6):1550–6.
18. Wass MN, Sternberg MJ. Confunc–functional annotation in the twilight zone. Bioinformatics. 2008;24(6):798–806.
19. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008;9(Suppl 1):4.
20. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol. 2010;8(2):357–76.
21. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins. 2011;79(7):2086–96.
22. Piovesan D, Tosatto SCE. INGA 2.0: improving protein function prediction for the dark proteome. Nucleic Acids Res. 2019;47(W1):373–8. https://doi.org/10.1093/nar/gkz375.
23. You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics. 2018;34(14):2465–73.
24. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multi-task deep neural networks. PLoS One. 2018;13(6):0198216.
25. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–7.
26. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184.
27. Friedberg I, Radivojac P. Community-wide evaluation of computational function prediction. Methods Mol Biol. 2017;1446:133–46.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
29. Kacsoh BZ, Barton S, Jiang Y, Zhou N, Mooney SD, Friedberg I, Radivojac P, Greene CS, Bosco G. New Drosophila long-term memory genes revealed by assessing computational function prediction methods. G3. 2019;9(1):251–67.
30. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. 2015;43(Database issue):1057–63.
31. Peng Y, Jiang Y, Radivojac P. Enumerating consistent sub-graphs of directed acyclic graphs: an insight into biomedical ontologies. Bioinformatics. 2018;34(13):313–22.
32. Wang L, Law J, Kale SD, Murali TM, Pandey G. Large-scale protein function prediction using heterogeneous ensembles. F1000Res. 2018;7.
33. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene CS. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst. 2017;5(1):63–71.
34. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45(Database issue):592–6.
35. Goyard S, Knechtle P, Chauvel M, Mallet A, Prevost MC, Proux C, Coppee JY, Schwarz P, Dromer F, Park H, Filler SG, Janbon G, d'Enfert C. The Yak1 kinase is involved in the initiation and maintenance of hyphal growth in Candida albicans. Mol Biol Cell. 2008;19(5):2251–66.
36. Gutierrez-Escribano P, Gonzalez-Novo A, Suarez MB, Li CR, Wang Y, de Aldana CR, Correa-Bordes J. CDK-dependent phosphorylation of Mob2 is essential for hyphal development in Candida albicans. Mol Biol Cell. 2011;22(14):2458–69.
37. Lassak T, Schneider E, Bussmann M, Kurtz D, Manak JR, Srikantha T, Soll DR, Ernst JF. Target specificity of the Candida albicans Efg1 regulator. Mol Microbiol. 2011;82(3):602–18.
38. Martin R, Moran GP, Jacobsen ID, Heyken A, Domey J, Sullivan DJ, Kurzai O, Hube B. The Candida albicans-specific gene EED1 encodes a key regulator of hyphal extension. PLoS One. 2011;6(4):18394.
39. Richard ML, Nobile CJ, Bruno VM, Mitchell AP. Candida albicans biofilm-defective mutants. Eukaryot Cell. 2005;4(8):1493–502.
40. Bernardo SM, Khalique Z, Kot J, Jones JK, Lee SA. Candida albicans VPS1 contributes to protease secretion, filamentation, and biofilm formation. Fungal Genet Biol. 2008;45(6):861–77.
41. Yi S, Sahni N, Daniels KJ, Lu KL, Huang G, Srikantha T, Soll DR. Self-induction of a/a or α/α biofilms in Candida albicans is a pheromone-based paracrine system requiring switching. Eukaryot Cell. 2011;10(6):753–60.
42. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, Paw J, Clore JJ, Mendoza RM, Luis BS, Nislow C, Giaever G, Costanzo M, Troyanskaya OG, Caudy AA. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLOS Genetics. 2009;5(3):1–16. https://doi.org/10.1371/journal.pgen.1000407.
43. Hibbs MA, Myers CL, Huttenhower C, Hess DC, Li K, Caudy AA, Troyanskaya OG. Directing experimental biology: a case study in mitochondrial biogenesis. PLOS Comput Biol. 2009;5(3):1–12. https://doi.org/10.1371/journal.pcbi.1000322.
44. Blus-Kadosh I, Zilka A, Yerushalmi G, Banin E. The effect of pstS and phoB on quorum sensing and swarming motility in Pseudomonas aeruginosa. PLoS One. 2013;8(9):74444.
45. Kuchma SL, Brothers KM, Merritt JH, Liberati NT, Ausubel FM, O'Toole GA. BifA, a cyclic-Di-GMP phosphodiesterase, inversely regulates biofilm formation and swarming motility by Pseudomonas aeruginosa PA14. J Bacteriol. 2007;189(22):8165–78.
46. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas Genome Database. Nucleic Acids Res. 2016;44(D1):646–53.
47. Friedman L, Kolter R. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. Mol Microbiol. 2004;51(3):675–90.
48. Friedman L, Kolter R. Two genetic loci produce distinct carbohydrate-rich structural components of the Pseudomonas aeruginosa biofilm matrix. J Bacteriol. 2004;186(14):4457–65.
49. Jackson KD, Starkey M, Kremer S, Parsek MR, Wozniak DJ. Identification of psl, a locus encoding a potential exopolysaccharide that is essential for Pseudomonas aeruginosa PAO1 biofilm formation. J Bacteriol. 2004;186(14):4466–75.
50. Synapse. https://www.synapse.org/. Accessed 1 Jan 2016.
51. Clark WT, Radivojac P. Information-theoretic evaluation of predicted ontological annotations. Bioinformatics. 2013;29(13):53–61.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

53. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):158–69.
54. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci USA. 2006;103(8):2833–8.
55. Noble SM, French S, Kohn LA, Chen V, Johnson AD. Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. Nat Genet. 2010;42(7): 590–8.
56. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Storms R, Boone C, Bussey H. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50(1):167–81.
57. Liu H, Kohler J, Fink GR. Suppression of hyphal formation in Candida albicans by mutation of a STE12 homolog. Science. 1994;266(5191): 1723–6.
58. You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S. NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res. 2019;47(W1):379–87. https://doi.org/10.1093/nar/gkz388.
59. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):447–52. https://doi.org/10.1093/nar/gku1003.
60. Dessimoz C, Skunca N, Thomas PD. CAFA and the open world of protein function predictions. Trends Genet. 2013;29(11):609–10.
61. Jiang Y, Clark WT, Friedberg I, Radivojac P. The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. Bioinformatics. 2014;30(17): 609–16.
62. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–83. https://doi.org/10.1093/bioinformatics/btg153. Accessed 1 Aug 2019.
63. Zhou N. Supplementary data. figshare. 2019. https://doi.org/10.6084/m9.figshare.8135393.v3. https://figshare.com/articles/Supplementary_data/8135393/3.
64. Jiang Y. CAFA2. Zenodo. 2019. https://doi.org/10.5281/zenodo.3403452.
65. Zhou N, Gerten M, Friedberg I. CAFA_assessment_tool. Zenodo. 2019. https://doi.org/10.5281/zenodo.3401694.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



A.11 Protein–protein interactions

Recommended reading for Chapter 11 (listed in Section 11.6).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Mering et al. (2002) — comparative assessment of large-scale interaction data; why different methods disagree.
  DOI · via SU library (SU login)

  Comparative assessment of large-scale data sets of protein–protein interactions. von Mering et al. (2002). *Nature*. doi:10.1038/nature750

  > Comprehensive protein protein interaction maps promise to reveal many aspects of the complex regulatory network underlying cellular function. Recently, large-scale approaches have predicted many new protein interactions in yeast. To measure their accuracy and potential as well as to identify biases, strengths and weaknesses, we compare the methods with each other and with a reference set of previously reported protein interactions.

- Luck et al. (2020) — HuRI, the reference map of the human binary interactome, and how its quality was assessed.
  DOI · via SU library (SU login)

  A reference map of the human binary protein interactome. Luck et al. (2020). *Nature*. doi:10.1038/s41586-020-2188-x

  > Global insights into cellular organization and genome function require comprehensive understanding of the interactome networks that mediate genotype-phenotype relationships 1,2 . Here we present a human 'all-by-all' reference interactome map of human binary protein interactions, or 'HuRI'. With approximately 53,000 protein-protein interactions, HuRI has approximately four times as many such interactions as there are high-quality curated interactions from small-scale studies. The integration of HuRI with genome 3 , transcriptome 4 and proteome 5 data enables cellular function to be studied within most physiological or pathological cellular contexts. We demonstrate the utility of HuRI in identifying the specific subcellular roles of protein-protein interactions. Inferred tissue-specific networks reveal general principles for the formation of cellular context-specific functions and elucidate potential molecular mechanisms that might underlie tissue-specific phenotypes of Mendelian diseases. HuRI is a systematic proteome-wide reference that links genomic variation to phenotypic outcomes.

- Barabási and Oltvai (2004) — network biology; the concepts of degree, hubs, modules and motifs.
  DOI · via SU library (SU login)

  Network biology: understanding the cell's functional organization. Barabási & Oltvai (2004). *Nature Reviews Genetics*. doi:10.1038/nrg1272

- Orth et al. (2010) — what flux balance analysis is, in a short primer.
  DOI · via SU library (SU login)

  What is flux balance analysis?. Orth et al. (2010). *Nature Biotechnology*. doi:10.1038/nbt.1614

- Bryant, Pozzati, and Elofsson (2022) — predicting protein–protein interactions and complex structures with AlphaFold2 (FoldDock, pDockQ).
  full text below · DOI · open access (CC BY 4.0)

  Improved prediction of protein-protein interactions using AlphaFold2. Bryant et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-28865-w

> Predicting the structure of interacting protein chains is a fundamental step towards understanding protein function. Unfortunately, no computational method can produce accurate structures of protein complexes. AlphaFold2, has shown unprecedented levels of accuracy in modelling single chain protein structures. Here, we apply AlphaFold2 for the prediction of heterodimeric protein complexes. We find that the AlphaFold2 protocol together with optimised multiple sequence alignments, generate models with acceptable quality (DockQ ≥ 0.23) for 63% of the dimers. From the predicted interfaces we create a simple function to predict the DockQ score which distinguishes acceptable from incorrect models as well as interacting from non-interacting proteins with state-of-art accuracy. We find that, using the predicted DockQ scores, we can identify 51% of all interacting pairs at 1% FPR.

- Burke et al. (2023) — towards a structurally resolved human protein interaction network.
  full text below · DOI · open access (CC BY 4.0)

  Towards a structurally resolved human protein interaction network. Burke et al. (2023). *Nature Structural & Molecular Biology*. doi:10.1038/s41594-022-00910-8

  > Cellular functions are governed by molecular machines that assemble through protein-protein interactions. Their atomic details are critical to studying their molecular mechanisms. However, fewer than 5% of hundreds of thousands of human protein interactions have been structurally characterized. Here we test the potential and limitations of recent progress in deep-learning methods using AlphaFold2 to predict structures for 65,484 human protein interactions. We show that experiments can orthogonally confirm higher-confidence models. We identify 3,137 high-confidence models, of which 1,371 have no homology to a known structure. We identify interface residues harboring disease mutations, suggesting potential mechanisms for pathogenic variants. Groups of interface phosphorylation sites show patterns of co-regulation across conditions, suggestive of coordinated tuning of multiple protein interactions as signaling responses. Finally, we provide examples of how the predicted binary complexes can be used to build larger assemblies helping to expand our understanding of human cell biology.

The open-access papers are reproduced below under their original licences.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28865-w OPEN

Improved prediction of protein-protein interactions using AlphaFold2

Patrick Bryant [1,2,3]✉, Gabriele Pozzati [1,2,3] & Arne Elofsson [1,2]✉

Predicting the structure of interacting protein chains is a fundamental step towards understanding protein function. Unfortunately, no computational method can produce accurate structures of protein complexes. AlphaFold2, has shown unprecedented levels of accuracy in modelling single chain protein structures. Here, we apply AlphaFold2 for the prediction of heterodimeric protein complexes. We find that the AlphaFold2 protocol together with optimised multiple sequence alignments, generate models with acceptable quality (DockQ ≥ 0.23) for 63% of the dimers. From the predicted interfaces we create a simple function to predict the DockQ score which distinguishes acceptable from incorrect models as well as interacting from non-interacting proteins with state-of-art accuracy. We find that, using the predicted DockQ scores, we can identify 51% of all interacting pairs at 1% FPR.

[1] Science for Life Laboratory, 172 21 Solna, Sweden. [2] Department of Biochemistry and Biophysics, Stockholm University, 106 91 Stockholm, Sweden. [3] These authors contributed equally: Patrick Bryant, Gabriele Pozzati. ✉email: patrick.bryant@scilifelab.se; arne@bioinfo.se

Protein–protein interactions are central mediators in biological processes. Most interactions are governed by the three-dimensional arrangement and the dynamics of the interacting proteins[1]. Such interactions vary from being permanent to transient[2,3]. Some protein–protein interactions are specific for a pair of proteins, while some proteins are promiscuous and interact with many partners. This complexity of interactions is a challenge both for experimental and computational methods.

Often, studies of protein–protein interactions can be divided into two categories, the identification of what proteins interact and the identification of how they interact. Although these problems are distinguished, some methods have been applied to both problems[4,5]. Protein docking methodologies refer to how proteins interact and can be divided into two categories considering proteins as rigid bodies; those based on an exhaustive search of the docking space[6] and those based on alignments (both sequence and structure) to structural templates[7]. Exhaustive approaches rely on generating all possible configurations between protein structures or models of the monomers[8,9] and selecting the correct docking through a scoring function, while template-based docking only needs suitable templates to identify a few likely candidates. However, flexibility has often to be considered in protein docking to account for interaction-induced structural rearrangements[10,11]. Therefore, flexibility limits the accuracy achievable by rigid-body docking[12], and flexible docking is traditionally too slow for large-scale applications. A possible compromise is represented by semi-flexible docking approaches[13] that are more computationally feasible and can consider flexibility to some degree during docking.

Regardless of different strategies, docking remains a challenging problem. In the CASP13-CAPRI experiments, human group predictors achieved up to 50% success rate (SR) for top-ranked docking solutions[14]. Alternatively, a recent benchmark study[8] reports SRs of different web-servers reaching up to 16% on the well-known Benchmark 5 dataset[15].

Recently, in the CASP14 experiment, AlphaFold2 (AF2) reached an unprecedented performance level in structure prediction of single-chain proteins[16]. Thanks to an advanced deep learning model that efficiently utilises evolutionary and structural information, this method consistently outperformed all competitors, reaching an average GDT_TS score of 90[16]. Recently, RoseTTAFold was developed, trying to implement similar principles[17]. Since then, other end-to-end structure predictors have emerged using different principles such as fast multiple sequence alignment (MSA) processing in DMPFold2[18] and language model representations[19].

As an alternative to other docking methods, it is possible to utilise co-evolution to predict the interaction between two protein chains. Initially, direct coupling analysis (DCA) was used to predict the interaction of bacterial two-component signalling proteins[20,21]. Later, these methods were improved using machine learning[22].

In a Fold and Dock approach, two proteins are folded and docked simultaneously. We recently developed a Fold and Dock pipeline using another distance prediction method focused on protein folding (trRosetta[23]). In this pipeline, the interaction between two chains from a heterodimeric protein complex and their structures were predicted using distance and angle constraints from trRosetta[24,25]. This study demonstrated that a pipeline focused on intra-chain structural feature extraction can be successfully extended to derive inter-chain features as well. Still, only 7% of the tested proteins were successfully folded and docked.

In that study, we found that generating the optimal MSA is crucial for obtaining accurate Fold and Dock solutions, but this is not always trivial due to the necessity to identify the exact set of interacting protein pairs[26]. Given the existence of multiple paralogs for most eukaryotic proteins, this is difficult. We also found that this process requires an optimal MSA depth to optimise inter-chain information extraction. Too deep MSAs might contain false positives (i.e. protein pairs that interact differently), resulting in noise masking the sought after co-evolutionary signal, while too shallow alignments do not provide sufficient co-evolutionary signals.

In this work, we systematically apply the AF2 pipeline on two different datasets to Fold and Dock protein–protein pairs simultaneously. We explore the docking success using the AF2 pipeline in combination with different input MSAs, in order to study the relationship between the output model quality and these inputs. We also find that, by scoring multiple models of the same protein–protein interaction with a predicted DockQ score (pDockQ), we can distinguish with high confidence acceptable (DockQ ≥ 0.23) from incorrect models. The modelling success is higher for bacterial protein pairs, pairs with large interaction areas consisting of helices or sheets, and many homologous sequences. We also test the possibility to distinguish interacting from non-interacting proteins and find that, using pDockQ, we can separate truly interacting from non-interacting proteins with consistent accuracy. We find that the results in terms of successful docking using AF2 are superior to other docking methods. AF2 clearly outperforms a recent state-of-the-art method[27] and our protocol performs quite close to (63% vs 72%) the recently developed AF-multimer[28], which was developed using the same data as the test set here, making a direct comparison difficult.

Results and discussion

Identifying the best AlphaFold2 model. The SR, i.e., the percentage of acceptable models (DockQ > 0.23), is used to measure AF2 performance over the development set (216 proteins) using the different MSAs. The best performance is 33.3% for the AF2 MSAs and 39.4% for the AF2+ paired MSAs (Table 1). It is thereby evident that combining both paired and AF2 MSAs is superior to using them separately. The average performance of the AF2 and the paired MSAs is similar, but for individual protein pairs, frequently one of the two MSAs is superior to the other, as seen from that the Pearson correlation coefficient for the DockQ scores between AF2 vs paired MSAs is 0.54 (Supplementary Table 1). Therefore, combining AF2 and paired MSAs improves the results.

Next, we compared the default AF2 model (model_1) with the fine-tuned versions of (model_1_ptm). Surprisingly, the original AF2 model_1 outperforms AF2 model_1_ptm in most cases (Table 1). Further, the difference between 10 recycles-one ensemble and three recycles-eight ensembles is minor across all MSAs and AF2 models. Therefore, the input information and the AF2 model appear to impact the outcome the most.

Table 1 Success rate of different modelling setups.

Neural network configuration				
NN model	model_1	model_1	model_1_ptm	model_1_ptm
Recycles	10	3	10	3
Ensembles	1	8	1	8
Setup short name	m1-10-1	m1-3-8	mp-10-1	mp-3-8
Paired MSA	**28.7**	28.2	**28.7**	27.8
AF2 MSA	31.5	**33.3**	26.4	23.6
AF2+Paired MSA	**39.4**	38.4	32.4	31.0

Results of AF2 run on the development set ($n = 216$) using different MSAs and neural network configurations. Row labels in bold indicate AF2 setup features. Every column in the table refers to an overall setup and every corresponding value refers to a run of the described setup with a different input MSA. Values represent the percentage of acceptable models (DockQ ≥ 0.23) overall the development set. The highest success rates for each MSA type are highlighted in bold.

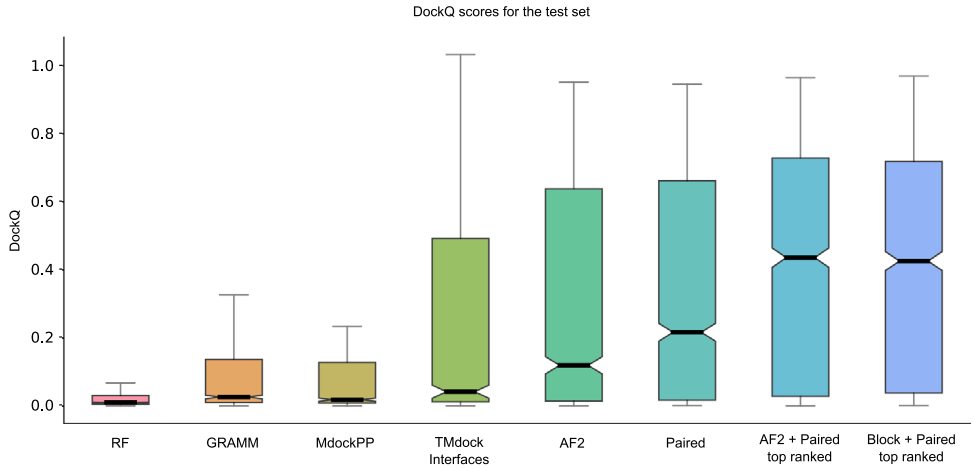


Fig. 1 DockQ scores for the test set ($n = 1481$ for all but RF, $n = 1455$). Distribution of DockQ scores as boxplots for different modelling strategies on the test set. Boxes encompass data quartiles, horizontal lines mark the medians and upper and lower whiskers indicate respectively maximum and minimum values for each distribution. All AF2 models have been run with the same neural network configuration (m1-10-1). Outlier points are not displayed here. AF2, refers to running AF2 using the default AF2 MSAs, "Paired" refers to using MSAs paired using information about species and "Block" refers to using block diagonalization MSAs.

Test set performance. The best model and configuration for AF2 (m1-10-1) was used for further studies on the test set. The best outcome using this modelling strategy results in an SR of 57.8% (856 out of 1481 correctly modelled complexes) for the AF2 + paired MSAs compared with 45.0% using the AF2 MSAs alone (Fig. 1, Table 2). The results using the block diagonalization+paired MSAs are almost identical (SR = 58.4%, median = 0.363). Further, running five initialisations with random seeds and ranking the models using the predicted DockQ score (pDockQ, Fig. 2c), increases the SR to 61.7% and 62.7% for the AF2 + paired and block diagonalization +paired MSAs, respectively (model variation and ranking, Fig. 2). Using the combination of AF2 and paired MSAs increases performance, suggesting that AF2 gains both from larger and paired MSAs, although it often can manage with less information.

What is most striking is that AF2 outperforms all other tested docking methods by a large margin (Fig. 1, Table 2). RF is better than AF2 only for 14 pairs in the test set, while GRAMM and template-based docking (TMdock interface) outperform AF2 for 188 and 225 pairs, respectively. The best performing method in the CASP14-CAPRI experiment[29], MDockPP[30], achieves a SR of only 24.2%. The reason GRAMM, TMdock and MDockPP reach this level of performance is likely due to the use of the bound form of the proteins, resulting in very high shape complementarity and therefore having the "answer" provided in a way.

The recently developed AF-multimer[28] has the best performance (SR = 72.2%, median = 0.560, Table 2). This method was trained using the same data as the test set, which makes a direct comparison difficult. Regardless, we do believe it is likely that using AF-multimer, the performance would increase over the results of our pipeline, but it is possible the difference is less than the observed 9%.

Table 2 Success rate and median DockQ scores for the test set using different methods and model configurations.

Method	Success rate (%)	Median DockQ
RoseTTAfold	9.6	0.011
GRAMM	21.4	0.027
MDockPP	24.2	0.019
TMdock	33.6	0.040
TMdock interfaces	35.1	0.042
AlphaFold2	45.0	0.120
Paired	49.6	0.217
AlphaFold2 + Paired	57.8	0.382
Block + Paired	58.4	0.363
AlphaFold2 + Paired top ranked	61.7	0.436
Block + Paired top ranked	62.7	0.426
AF-multimer	72.2%	0.560

"Block" refers to block diagonalization MSAs. Results for different docking methods in terms of success rate and median DockQ scores on the test set. The number of complexes is $n = 1481$ for all methods except for RoseTTAfold ($n = 1455$) and AF-multimer ($n = 1458$).

Distinguishing acceptable from incorrect models. It is not only essential to obtain improved predictions, but also to be able to discriminate between acceptable and non-acceptable ones. We measure the separation between correct (DockQ ≥ 0.23) and incorrect models provided by several metrics using a receiver operating characteristic (ROC) curve. Different criteria were examined over the test set, including (i) the number of unique interacting residues (Cβ atoms from different chains within 8 Å from each other) in the interface, (ii) the total number of interactions between Cβ atoms in the interface, (iii) the average plDDT for the interface, (iv) the lowest plDDT of each single-chain average, and (v) the average plDDT over the whole protein heterodimer (Fig. 2a). Three criteria result in very similar areas under the curve (AUC) measures. The total number of interactions between Cβs and the number of residues in the interface can separate the correct/incorrect models with an AUC of 0.92 and 0.91 respectively, while the average interface plDDT results in an AUC of 0.88. However, pLDDT results in higher TPRs at lower FPRs; therefore, we multiply the plDDT with the logarithm of the interface contacts resulting in an AUC of 0.95.

Interestingly, the average plDDT of the entire complex only results in an AUC of 0.66, suggesting that both single chains in a complex are often predicted very well, while their relative orientation may still be incorrect.

Figure 2b shows that increasing both the number of interface contacts and the average interface plDDT results in higher DockQ scores for the test set. Using the combination of plDDT with the logarithm of the interface contacts, we, therefore, fit a simple sigmoidal function to the DockQ scores (Fig. 2c), see methods. This enables the prediction of the DockQ scores

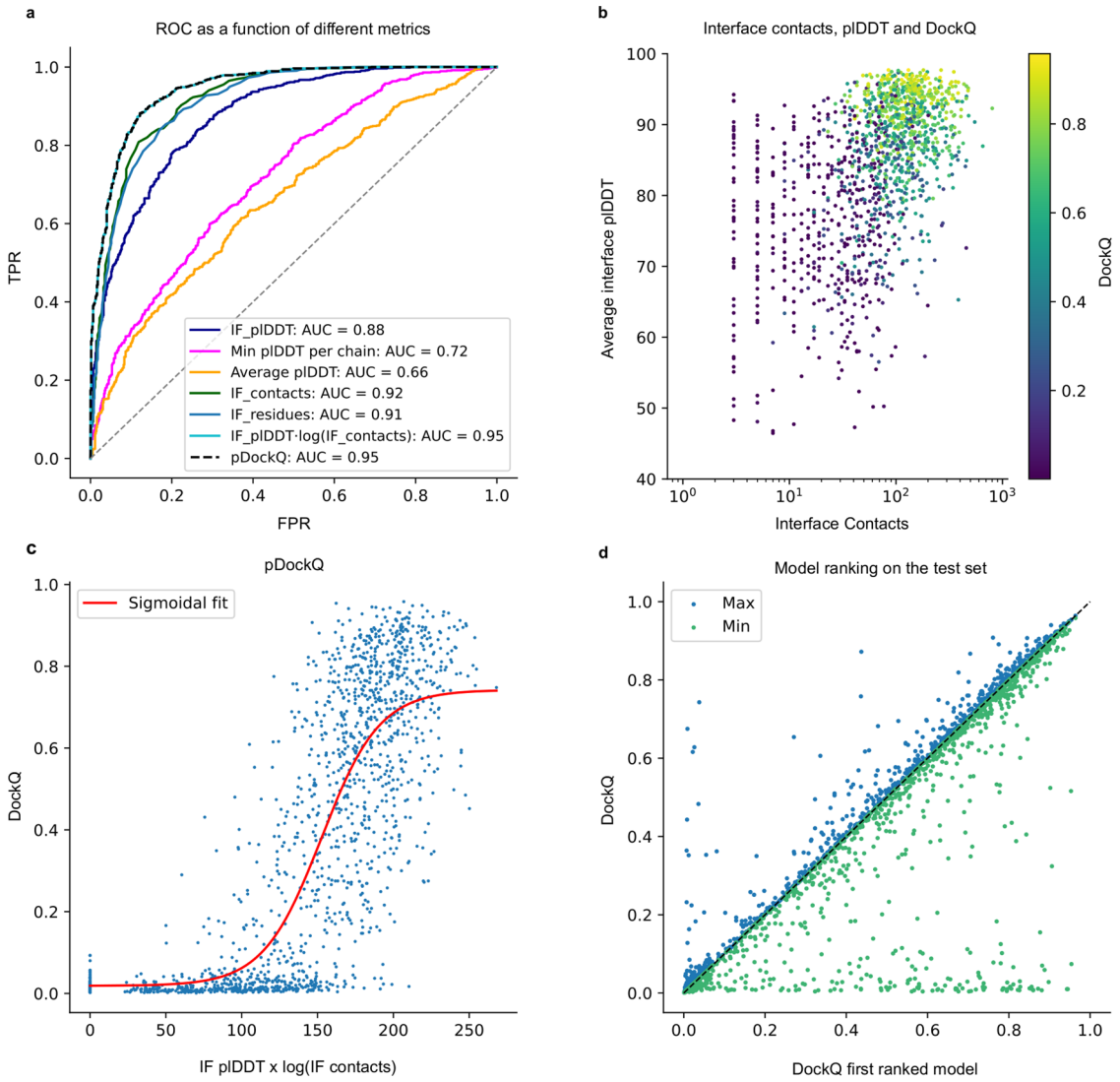


Fig. 2 Model quality metrics and multiple model ranking. a ROC curve as a function of different metrics for the test dataset ($n = 1481$, first run). Cβs within 8 Å from each other from different chains are used to define the interface. IF_plDDT is the average plDDT of interface residues, min plDDT per chain is the minimum average plDDT of both chains, average plDDT is the average of the entire complex and IF_contacts and IF_residues are the number of interface residues and contacts respectively. pDockQ is a sigmoidal fit to the combined metric IF_plDDT·log(IF_contacts) fitted to predict DockQ as the target score, see C. **b** Average interface plDDT vs the logarithm of the interface contacts coloured by DockQ score on the test set ($n = 1481$). Increasing both the number of interface contacts and average interface plDDT results in higher DockQ scores. **c** Using the combined metric IF_plDDT·log(IF_contacts), we fit a sigmoidal curve towards the DockQ scores on the test set ($n = 1481$), enabling predicting the DockQ score in a continuous manner (pDockQ). The average error overall is 0.14 DockQ score. **d** Impact of different initialisations on the modelling outcome in terms of DockQ score on the test dataset ($n = 1481$). The maximal and minimal scores are plotted against the top-ranked models using the pDockQ scores for the AF2 + paired MSAs, m1-10-1.

(pDockQ) in a continuous manner with an overall average error of 0.11 on the test set. The AUC using pDockQ as a separator is identical to the combination of plDDT with the logarithm of the interface contacts, 0.95 (Fig. 2a).

Model variation and ranking for the test set. Five models are generated using the best strategy (m1-10-1 with AF2 + paired MSAs) with different initialisation (random seeds). The average SR (57.2% ± 0.0%) is similar for all five runs. However, the average deviation for individual models is DockQ = 0.08 when comparing the best and worst models for a target (Fig. 2d), i.e., there is some randomness to the success for an individual pair. If the maximal DockQ score across all models is used, the SR would be 62.9%. Although this is unachievable, ranking the models using the pDockQ score results in an SR of 61.7%. The AUC using the same metric for the ranked test set is 0.93, which means that 31% of all models are acceptable at an error rate of 1% and 54% at an error rate of 10% (Supplementary Table 2).

Bacterial complexes are predicted more accurately. In the test set, about 60% of the complexes can be modelled correctly. We try to identify what distinguishes the successful and unsuccessful cases by analysing different subsets of the test set. First, we divide the proteins by taxa, next by interface characteristics and finally by examining the alignments.

The SRs for each kingdom is; Eukarya 61%, Bacteria 73.7%, Archaea 84.5%, and Virus 60% (Supplementary Fig. 1b). Further, the SRs for *Saccharomyces cerevisiae* is better than for *Homo sapiens* (66% vs 58%, Fig. 3d). The higher performance in prokaryotes is consistent with previous observations regarding the availability of evolutionary information in prokaryotes compared to Eukarya[27] (Supplementary Fig. 1). The higher

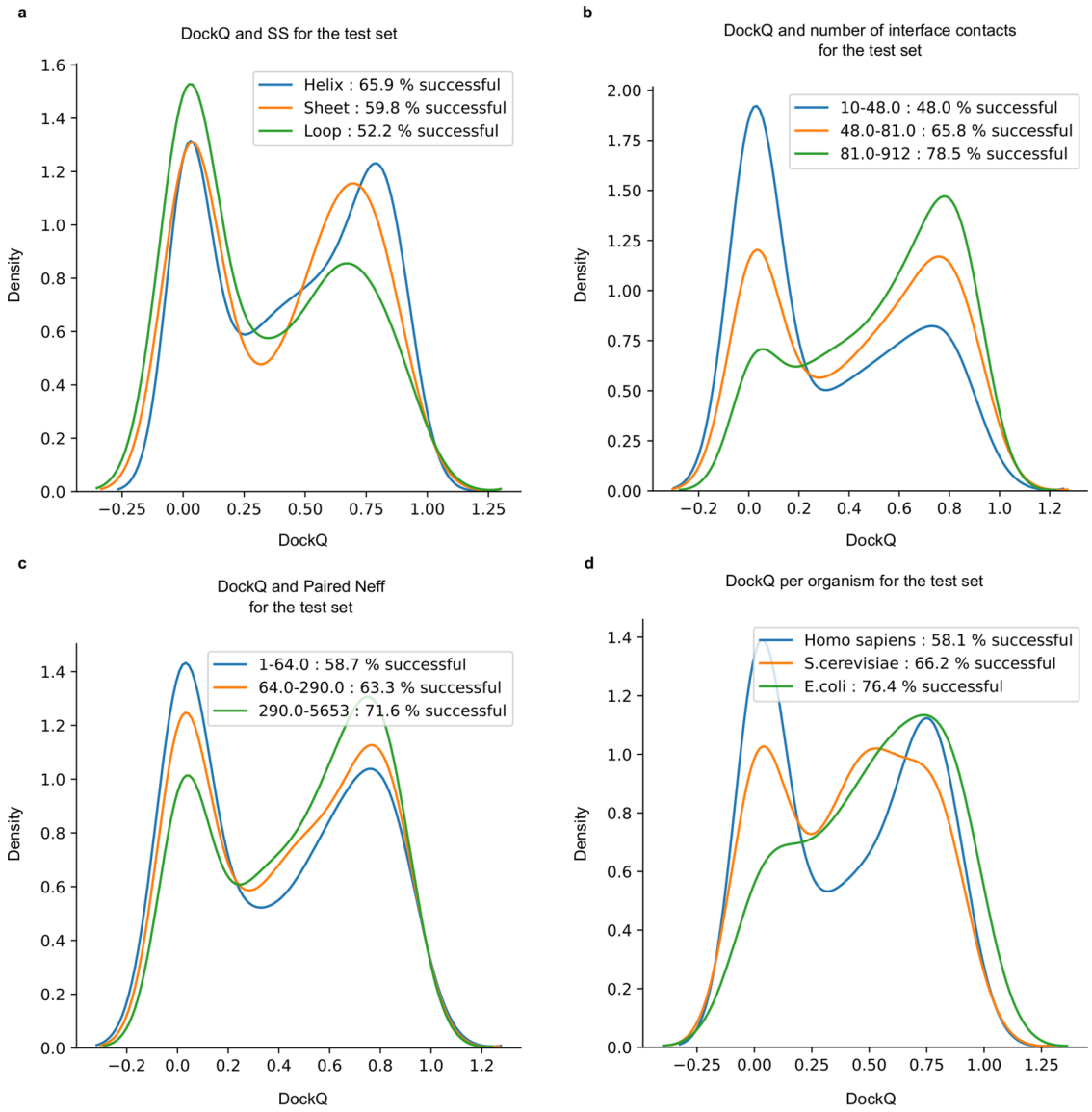


Fig. 3 DockQ distributions for test dataset ($n = 1481$) tertiles. a Distribution of DockQ scores for three sets of interfaces with the majority of Helix, Sheet and Coil secondary structures. **b** Distribution of DockQ scores for tertiles derived from the distribution of contact counts in docking model interfaces. **c** Distribution of DockQ scores for tertiles derived from the distribution of Paired MSAs Neff scores. **d** Distribution of DockQ scores for the top three organisms *H. sapiens*, *S. cerevisiae* and *E. coli*.

performance in *S. cerevisiae* compared to *H. sapiens* suggests a similar relationship between higher and lower order organisms within the same kingdom.

Next, we examine the interfaces. Different secondary structural content of the native interfaces is investigated (Fig. 3a). The highest SR is obtained mainly for helix interfaces (62%), followed by interfaces containing mainly sheets (59%). The loop interface SR of 53% is substantially lower than the others, suggesting that interfaces with more flexible structures are harder to predict. We divide the dataset by interface size, and find that pairs with larger interfaces are easier to predict, as the SR increases from 47 to 74% between the smallest and biggest tertiles (Fig. 3b).

We continue to examine features of the MSAs. First, the impact of the number of non-redundant sequences (Neff) in both paired and AF2 MSAs was analysed. It is clear that the fraction of correctly modelled sequences increases with larger Neff scores (Fig. 3c). Also, paired MSA Neff (Fig. 3c) has a stronger influence on the outcome than the Neff of the AF2 MSAs (Supplementary Fig. 2a). Secondly, the MSA interface signal in the paired MSAs, measured by the fraction of correct interface contacts using DCA, was analysed. MSAs with stronger interface signals show higher SRs, even if the paired MSAs are used in combination with the AF2 MSAs (Supplementary Fig. 3). This suggests that MSA co-evolutionary signal and, thereby, correct identification of orthologous protein sequences, has a strong impact on the outcome.

CASP14 and novel proteins without templates. Chains derived from CASP14 heteromeric targets and chains from PDB complexes with no templates are folded in pairs using the presented AF2 pipeline (default AF2 + paired MSAs, ten recycles, m1-10-1 and five differently seeded runs).

For the CASP14 chains, four out of six pairs display a DockQ score larger than 0.23 (SR of 67%). No ranking is necessary in this case, given that all produced docking models for the same chain pair are very similar (the average standard deviation is 0.01 between each set of DockQ scores). An interesting unsuccessful docking is obtained modelling chains from the complex with PDB

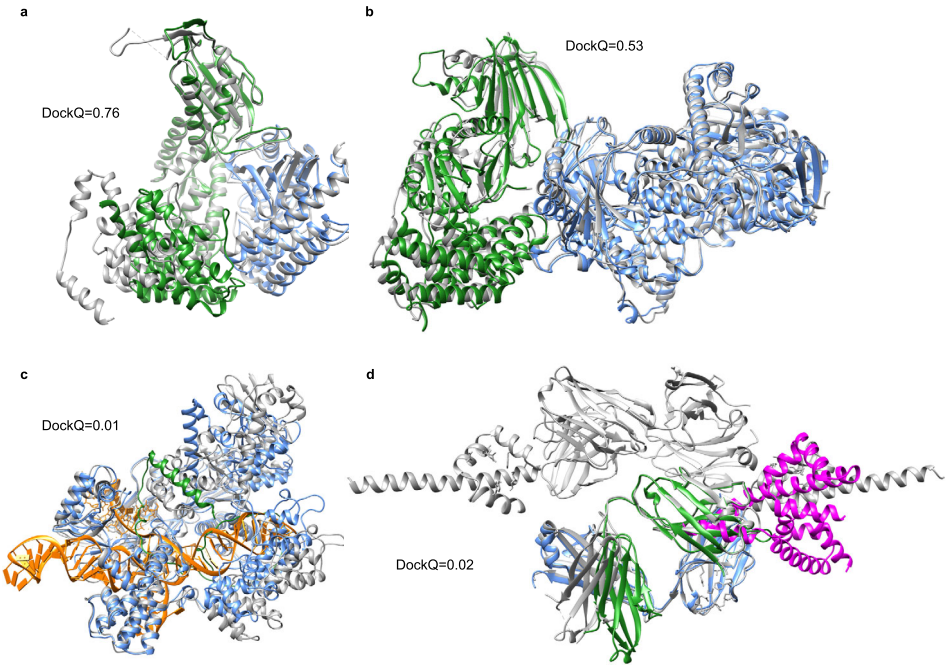


Fig. 4 Predicted and native structures from the set of novel proteins without templates. The native structures are represented as grey ribbons. **a** Docking of 7EIV chains A (blue) and C (green) (DockQ = 0.76). **b** Docking of 7MEZ chains A (blue) and B (green) (DockQ = 0.53). **c** Prediction of structure 7EL1 chains A (blue) and E (green) (DockQ = 0.01). The DNA going through chain A is coloured in orange. **d** Docking of 7LF7 chains A (blue) and M (magenta) (DockQ = 0.02) and chains B (green) and M (magenta) (DockQ = 0.02).

ID 6TMM (Supplementary Fig. 4), which are known to form a heterotetramer. In this structure, each chain A is in contact with its partner chain B at two different sites. Both docking configurations (6TMM_A-B and 6TMM_A-D) put the chain in between the two binding sites. The other unsuccessful docking (6VN1_A-H) has an interface of just 19 residue pairs.

The SR for docking the proteins without templates is 50%. Between the five different initialisations, the average difference in the DockQ score is 0.03, and there is no deviation in SR, i.e., ranking did not improve the SR. Two acceptable models are displayed in Fig. 5a (7EIV_A-C]) and B (7MEZ_A-B). More interesting, in one of the incorrect models (7NJ0_A-C], Supplementary Fig. 5), the predictions get the location of both chains correct, but their orientations wrong, resulting in DockQ scores close to 0. For 7EL1_A-E (Fig. 4c), the shorter chain E is not folded correctly, and instead of folding to a defined shape, it is stretched out and inserted within chain A. It occupies the shape of the DNA in the native structure. In the two remaining incorrect models (7LF7_A-M and 7LF7_B-M), Fig. 4d, the chains only interact with a short loop of the M chain, making the docking very difficult and possibly biologically meaningless.

Identifying interacting proteins. Using the best separator from the model ranking, the pDockQ, it is possible to distinguish the 3989 non-interacting proteins from *Escherichia coli* and the 1481 truly interacting proteins from the test set with an AUC of 0.87. Another recently published method obtains AUC 0.76 on this set[27]. However, these results are probably overstated since the negative set only contains bacterial proteins, while the positive set is mainly eukaryotic.

To obtain a more realistic estimate, we also include a set of 1705 non-interacting proteins from mammalian organisms[31] combined with the non-interacting proteins from *E. coli*. On this combined set of 1481 interacting and 5694 non-interacting proteins, we obtain an AUC of 0.82 for the average interface plDDT and slightly higher (0.84 and 0.85) for the number of interface contacts and residues, respectively (Fig. 5a). pDockQ results in an ROC curve with an AUC of 0.87. Importantly, pDockQ provides a better separation at low FPRs, enabling a TPR of 51% at FPR of 1% compared to 27%, 18 and 13% for the interface plDDT, number of interface contacts and residues, respectively. At FPR 5%, the number of interface contacts and residues report TPRs of 49 and 42%, respectively, compared to 43% for the average interface plDDT and 66% for pDockQ. The distribution of the top separators can be seen in Fig. 5b–d.

Limitations. Here, we only consider the structures of protein complexes in their heterodimeric state, although each protein chain in these complexes may have homodimer configurations or other higher-order states. It is also possible that the complex itself exists as part of larger biological units, in potentially more complex conformations. Investigating alternative oligomeric states and larger biological assemblies is outside of the scope of this analysis and left for future work.

The study of AF2s ability to separate interacting and non-interacting proteins here contains more extensive data than recent studies[27]. However, to test this separation thoroughly, the data studied here needs to be extended to compare interactions within individual organisms. We leave this extensive analysis to further studies.

There is a big difference between the performance of AF2 on the development and test sets, reporting 39.4% SR vs 57.8% for the AF2 + Paired MSAs. This discrepancy suggests that the performance is highly dependent on the specific interacting partners being predicted. It is not clear what causes this difference as the composition in terms of kingdom, found to be very important (Supplementary Fig. 1b) is similar (54% vs 60% Eukaryotic proteins), the MSAs have similar Neff scores (2699 vs. 2764 on average), the proteins are of similar sizes (222 vs. 203 AAs on average), and the number of residues in the interface is similar (139 vs 120 on average). This leads us to believe that there may be some unknown selection bias in how the sets were chosen. It can be noted that the development is much smaller than the

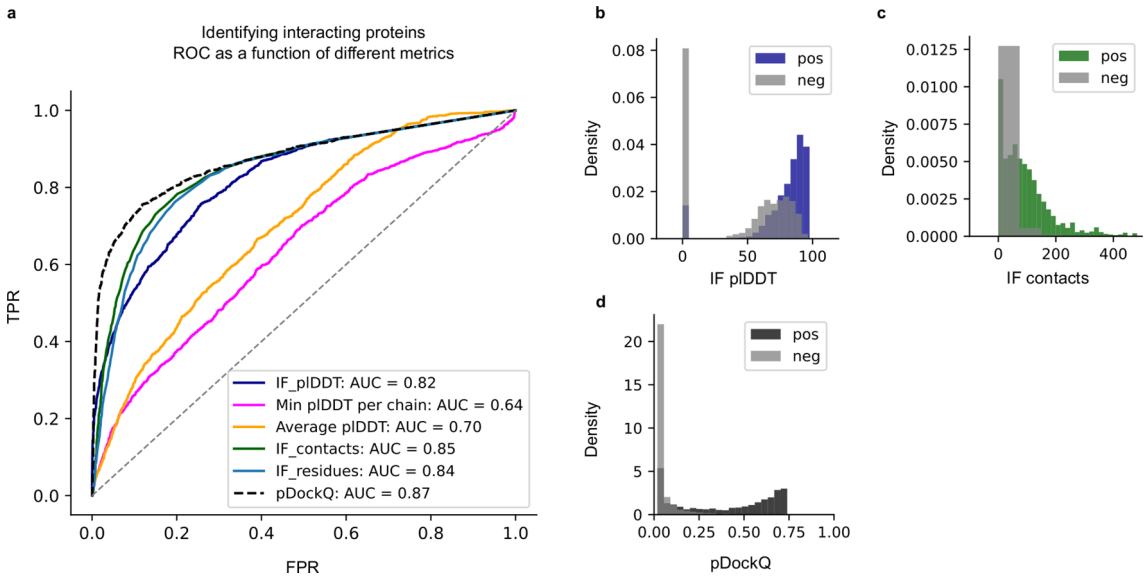


Fig. 5 Discrimination of interacting ($n = 1481$) and non-interacting ($n = 5694$) proteins. a The ROC curve as a function of different metrics for discriminating between interacting and non-interacting proteins. IF_plDDT is the average plDDT in the interface, min plDDT per chain is the minimum average plDDT of both chains, average plDDT is the average of the entire complex and IF_contacts and IF_residues are the number of interface residues and contacts respectively. pDockQ is a sigmoidal fit to this with DockQ as the target score, as described above. **b**–**d** Distribution of the top discriminating features average interface plDDT (**b**), the number of interface contacts (**c**), and **d** the combination of these (IF_plDDT·log(IF_contacts)) and the pDockQ for interacting (non-grey) and non-interacting proteins (grey).

test set though (216 vs 1481 proteins), which is why performance should be assessed on as large non-redundant datasets as possible.

Findings and future prospects. Here we show that AlphaFold2[16] (AF2) can predict the structure of many heterodimeric protein complexes, although it is trained to predict the structure of individual protein chains. Even using the default settings, it is clear that AF2 is superior to all other tested docking methods, including other Fold and Dock methods[17,24], methods based on shape complementarity[30,32] and template-based docking. Using optimised MSAs with AF2, we can accurately predict the structure of heterodimeric complexes for an unprecedented SR of 62.7% on a large test set. The SR is higher in *E.coli* (76.4%) than in *H. sapiens* or *S. cerevisiae* (58.1% and 66.2% respectively).

Further, by analysing the predicted interfaces, we can predict the DockQ score[33] (pDockQ) with an average error of 0.1, resulting in the separation of acceptable and incorrect models with an AUC of 0.95. This means that 31% of the models can be called acceptable at a specificity of 99% (or 54% at 90% specificity). Interestingly, no additional constraints are implemented in AF2 to pull two chains in contact, meaning that chain interactions (and subsequently interface sizes) are exclusively determined by the amount of inter-chain signals extracted by the predictor. Assuming that all residues in an interface contribute to the interaction energy could explain why larger interfaces are more likely to be correctly predicted.

We find that the MSA generation process can be sped up substantially at no performance loss (performance increase of 1% SR) by simply fusing MSAs from two HHblits[34] runs on Uniclust30[35] instead of using the MSAs from AF2. Fast MSA generation circumvents the main computational bottleneck in the pipeline. Using pDockQ makes it possible to separate truly interacting from non-interacting proteins with an AUC of 0.87, making it possible to identify 51% of interacting proteins at an error rate of 1%. The pDockQ score discriminates between both model quality and binary interactions. Therefore, the same pipeline can identify if two proteins interact and the accuracy of their structure.

Never before has the potential for expanding the known structural understanding of protein interactions been this large, at such a small cost. There are currently 64,006 pairwise human protein interactions in the human reference interactome[36]. If 31% of these can be predicted at an error rate of 1%, this results in the structure of 19,842 human heterodimeric protein structures. The computational cost to run all of this is ~5 days on an Nvidia A100 system and has since the development of the pipeline presented here, deemed FoldDock, been applied[37].

Methods

Development set. A set of heterodimeric complexes from Dockground benchmark 4[38] is used to develop the pipeline, focusing on the AF2 configuration presented here. This set contains protein pairs, with each chain having at least 50 residues, sharing <30% sequence identity and no crystal packing artefacts. There are 219 protein interactions for which both unbound (single-chain) and bound (interacting chains) structures are available. Unbound chains share at least 97% sequence identity with the bound counterpart and, to facilitate comparisons, non-matching residues are deleted and renumbered to become identical to the unbound counterpart. AF2 MSAs could not be generated for three of the complexes due to memory limitations (1gg2, 2nqd and 2xwb) using a computational node with 128 Gb RAM for the MSA generation and were thus disregarded, resulting in a total of 216 complexes. The dataset consists of 54% Eukaryotic proteins, 38% Bacterial and 8% from mixed kingdoms, e.g., one bacterial protein interacting with one eukaryotic.

Test set. We used 1661 protein complexes with known interfaces from a recent study[27] to test the developed pipeline. Here, three large biological assemblies were excluded. These complexes share <30% sequence identity, have a resolution between 1–5 Å and constitute unique pairs of PFAM domains (no single protein pair have PFAM domains matching that of any other pair). Some structures failed to be modelled for various reasons (see limitations of data generation), resulting in a total of 1481 structures. These proteins are mainly from *H. sapiens* (25%), *S. cerevisiae* (10%), *E. coli* (5%) and other Eukarya (30%).

107 of the complexes in the test set lack beta carbons (Cβs), and 50 have overlapping PDB codes with the development set and were therefore excluded. In the MSA generation from AF2, 20 MSAs report MergeMasterSlave errors regarding discrepancies in the number of match states, resulting in a total of 1484 AF2 MSAs. When folding, three of these (5AWF_D-5AWF_B, 2ZXE_B-2ZXE_A and 2ZXE_A-2ZXE_G) report "ValueError: Cannot create a tensor proto whose content is larger than 2GB", leading to a final set of 1481 complexes. DSSP could only be run successfully for 1391 out of the 1481 protein complexes, and we ignored the rest in the analysis.

For RF, 26 complexes produced out of memory exceptions during prediction using a GPU with 40 Gb RAM and were excluded from the RF analyses, leaving 1455 complexes.

For the mammalian proteins from Negatome, seven out of 1733 single chains were redundant according to Uniprot (C4ZQ83, I0LJR4, I0LL25, K4CRX6, P62988, Q8NI70, Q8T3B2), 34 had no matching species in the MSA pairing, 106 produced out of memory exceptions during prediction using a GPU with 40 Gb RAM, 35 gave a tensor reshape error, and 65 complexes were homodimers, leaving 1715 complexes for this set.

CASP14 set and novel protein complexes. As an additional test set, we used a set of six heterodimers from the CASP14 experiment. In addition, we extracted eight novel protein complexes deposited in PDB after 15 June 2021, which produced no results for at least one chain in each complex when submitted to the HHPRED web server (version 01-09-2021)[39,40], see Supplementary Table 3. We selected this small set to test the performance on data AF2 is guaranteed not to have seen.

Non-interacting proteins. Two datasets of known non-interacting proteins were used, one from the same study as the positive test set[27]. Here, all proteins are from *E. coli*. Two methods were used to identify non-interacting proteins, first a set of proteins with no reported interaction signal in Yeast Two-Hybrid Experiments[41] and secondly complexes whose individual proteins were found in different APMS benchmark complexes[42]. This dataset contains in total 3989 non-interacting pairs.

The second set contains 1964 unique mammalian protein complexes filtered against the IntAct[43] dataset from Negatome[31]. This data deemed "the manual stringent set" contains proteins annotated from the literature with experimental support describing the lack of protein interaction. Some structures in this dataset are homodimers (65) and are therefore excluded, resulting in 1705 structures. Together there are 5694 non-interacting protein complexes.

AlphaFold2 default MSA generation methodology. The input to AlphaFold2 (AF2) consists of several MSAs. We used the AF2 MSA generation[16], which builds three different MSAs generated by searching the Big Fantastic Database[44] (BFD) with HHBlits[34] (from hh-suite v.3.0-beta.3 version 14/07/2017) and both MGnify v.2018_12[45] and Uniref90 v.2020_01[46] with jackhmmer from HMMER3[47]. The AF2 MSAs were generated by supplying a concatenated protein sequence of the entire complex to the AF2 MSA generating pipeline in FASTA format. The resulting MSAs will thus mainly contain gaps for one of the two query proteins in each row, as only single chains can obtain hits in the searched databases (Fig. 6). No trimming or gap removal was performed on these MSAs.

MSA block diagonalization. In addition to the default AF2 MSA, we generated an additional MSA by simply concatenating diagonally MSAs generated independently from each of the two chains. These MSAs were constructed by running HHblits[34] version 3.1.0 against uniclust30_2018_08[35] with these options:

```
hhblits -E 0.001 -all -oa3m -n 2
```

The concatenation is done by joining side-by-side the two input chains; then sequences from one MSA are added, aligned to the corresponding input chain. Each sequence in the MSA is then elongated with gaps (on the right side if it is the left sequence MSA or the other way around), to reach the length of the two concatenated input chains. The process is then repeated for the other input chain MSA to complete the block diagonalization.

Paired MSA generation. In addition to the block diagonalization MSAs, we used a "paired MSA", constructed using organism information, where sequences are matched based on their organism origins[4,21,24] (Fig. 6). The rationale behind using a paired MSA is to identify inter-chain co-evolutionary information. An unpaired MSA has a limited inter-chain signal since the chains are treated in isolation.

The organism information was, using the OX identifier, extracted from the two HHblits MSAs[48]. Next, all hits with more than 90% gaps were removed. From all remaining hits in the two MSAs, the highest-ranked hit from one organism was paired with the highest-ranked hit of the interacting chain from the same organism. Pairing the correct sequences should result in MSAs containing inter-chain co-evolutionary information[27].

Number of effective sequences (Neff). To estimate the information in each MSA, we clustered sequences at 62% identity, as described in a previous study[50]. The number of clusters obtained in this way has been used to indicate a N_{eff} value for each MSA.

Unaligned FASTA sequences were extracted from the three AF2 default MSAs. Obtained sequences were processed with the CD-HIT software[51] version 4.7 (http://weizhong-lab.ucsd.edu/cd-hit/) using the options:

```
-c 0.62 -G 0 -n 3 -aS 0.9
```

We calculated the Neff scores separately for paired and AF2 MSAs.

AlphaFold2. We modelled complexes using AlphaFold2[16] (AF2) by modifying the script https://github.com/deepmind/alphafold/blob/main/run_alphafold.py to insert a chain break of 200 residues—as suggested in the development of RoseTTAFold[17] (RF). During modelling, relaxation was turned off. We note that performing model relaxation did not increase performance in the AF2 paper[16] and was, therefore, ignored to save computational cost. No templates were used to build structures, as this would not assess the prediction accuracy of unknown structures or structures without sufficient matching templates. Further, AF2 has been shown to perform well for single chains without templates and has reported higher accuracy than template-based methods even when robust templates are available[16].

We supplied four different types of MSAs to AF2: (1) the MSAs generated by using the default AF2 settings, (2) the top paired MSAs constructed using HHblits, described above, (3) both alignments together and finally, (4) the top paired and single-chain MSAs from HHblits to speed up predictions (only for the test set). AF2 was run with two different network models, AF2 model_1 (used in CASP14) and AF2 model_1_ptm, for each MSA. The second model, model_1_ptm, is a fine-tuned version of model_1 that predicts the TMscore[52] and alignment errors[16]. We ran these two different models by using two different configurations. The configurations utilise a varying amount of recycles and ensemble structures. Recycle refers to the number of times iterative refinement is applied by feeding the intermediate outputs recursively back into the same neural network modules. At each recycling, the MSAs are resampled, allowing for new information to be passed through the network. The number of ensembles refers to how many times information is passed through the neural network before it is averaged[16]. The two configurations used are; the CASP14 configuration (three recycles, eight ensembles) and an increased number of recycles (ten) but only one ensembles.

Since structure prediction with AF2 is a non-deterministic process, we generate five models initiated with different seeds. To save computational cost, this was only performed for the best modelling strategy. We rank the five models for each complex by the number of residues in the interface, giving the best result.

RoseTTAFold. For comparison, the RoseTTAFold (RF) end-to-end version[17] was run using the paired MSAs with the top hits. The RoseTTAFold pipeline for complex modelling only generates MSAs for bacterial protein complexes, while the proteins in our test set are mainly Eukaryotic. Therefore, we use the paired alignments here. We compare RF with AF2 using the same inputs (the paired MSAs) for both the development and test datasets to provide a more fair comparison, as AF2 searches many different databases to obtain as much evolutionary information as possible when generating its MSAs. To predict the complexes, we use the "chain break modelling" as suggested in RF (https://github.com/RosettaCommons/RoseTTAFold/tree/main/example/complex_modeling) using the following command:

```
predict_complex.py -i msa.a3m -o complex -Ls chain1_length
chain2_length
```

No optimisation of the RF protocol was made here.

MDockPP. The docking method MDockPP[30] was run through the provided webserver (https://zougrouptoolkit.missouri.edu/MDockPP/). This docking algorithm is based on fast Fourier transform (FFT). The docking results are assessed using the "in-house" scoring function ITScorePP.

GRAMM. For comparison, a rigid-body docking method, GRAMM[32], was used. Here, two protein models are docked using a FFT procedure to generate 340,000 docking poses for each complex. The bound structures extracted from complexes in the test set were used as inputs. This docking generation stage mainly considers the geometric surface properties of the two interacting structures, allowing minor clashes to leave some space for conformational flexibility adjustment. As the bound form of the proteins is used, this should represent an easy case for GRAMM-based docking, and performance drops significantly when unbound structures or models are used[53]. The atom-atom contact energy AACE18 is used to score and rank all poses, as this has been shown to provide better results than shape-complementarity alone[54].

Template-based docking. For comparison, a template-based docking protocol[7] referred to as "TMdock" is also adopted. The adopted template library includes 11756 protein complexes obtained from the Dockground database[38] (release 28-10-2020). Monomers from target complexes are structurally aligned with complexes in the supplied libraries (depleted of the target structure PDB ID) in order to identify the best available template structure. The bound form of the template structures was used. TM scores resulting from the alignment of target proteins to each template are averaged and used to score obtained docking models. Alternatively, we refer to "TMdock Interfaces" when targets are structurally aligned only to the template interfaces, defined as every residue with a Cβ atom closer than 12 Å from any Cβ atom in the other chain.

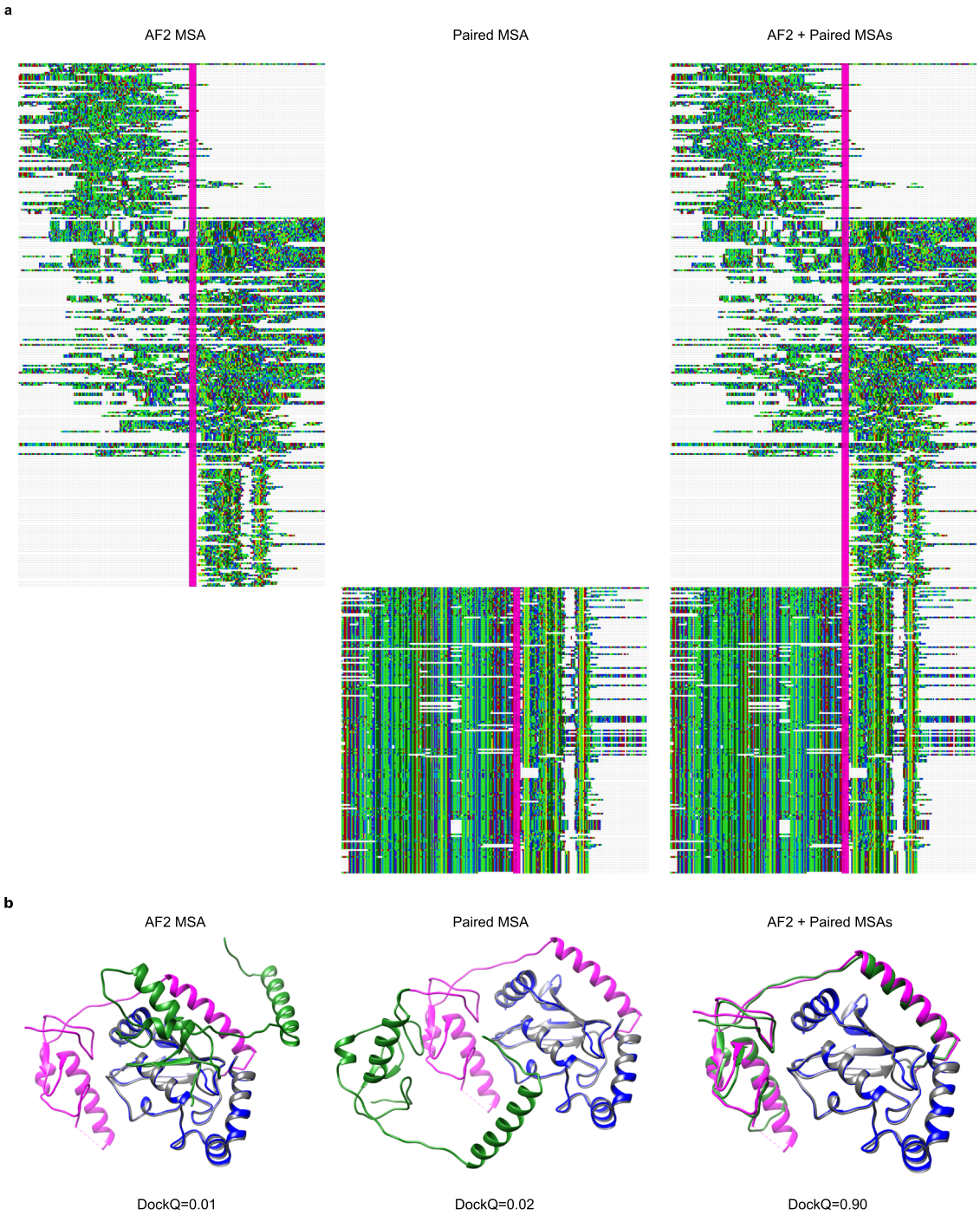


Fig. 6 Comparison of different MSAs. a Depiction of MSAs generated by AF2 and the paired version matched using organism information. Both AF and paired representations are sections containing 10% of the sequences aligned in the original MSA. Concatenated chains are separated by a vertical line (magenta). The visualisations were made using Jalview version 2.11.1.4[49]. **b** Docking visualisations for PDB ID 5D1M with the model/native chains A in blue/grey and B in green/magenta using the three different MSAs in (**a**). The DockQ scores are 0.01, 0.02 and 0.90 for AF2, paired, and AF2 + paired MSAs, respectively.

AlphaFold-multimer. The simultaneous fold-and-dock program based on the same principles as AF2, AlphaFold-multimer[28], was run with the default settings. These entail creating four different MSAs. Three different MSAs are created by searching Uniref90 v.2020_01[46], Uniprot v.2021_04[48] and MGnify v.2018_12[45] with jackhmmer from HMMER3[47] and one joint is created by searching the Big Fantastic Database[44] (BFD) and uniclust30_2018_08[35] with HHBlits[34] (from hh-suite v.3.0-beta.3 version 14/07/2017).

The results from the Uniprot search are used for MSA pairing and the results from the remaining searches are used to create a block-diagonalized MSA, similar to the procedures described above. All four MSAs are then used to fold a protein complex. Some complexes failed due to computational limitations, resulting in 1458 out of 1481 complexes successfully folded.

Scoring models. The backbone atoms (N, CA and C) were extracted from the predicted AF2 structures (as these are the only predicted atoms in the end-to-end version of RF). The interface scoring program DockQ[33] was then run (without any special settings) to compare the predicted and actual interfaces. This program

compares interfaces using a combination of three different CAPRI[55] quality measures (F_{nat}, LRMS, and iRMS) converted to a continuous scale, where an acceptable model comprises a DockQ score of at least 0.23.

Ranking models. To analyse the ability of AF2 to distinguish correct models as well as interacting from non-interacting proteins, we analyse the separation between acceptable and incorrect models as a function of different metrics on the development set: the number of unique interacting residues (Cβs from different chains within 8 Å from each other), the total number of interactions between Cβs from different chains (referred to as the number of interface contacts), average predicted lDDT (plDDT) score from AF2 for the interface, the minimum of the average plDDT for both chains and the average plDDT over the whole heterodimer.

We use these metrics as a threshold to build a confusion matrix, where true/false positives (TP and FP respectively) are correct/incorrect docking models which places above the threshold and false/true negatives (FN and TN respectively) are correct/incorrect docking models which scores below the threshold. From the built confusion matrix, we derive the true positive rate (TPR), false positive rate (FPR) defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2}$$

Then, we calculate TPR and FPR for each possible value assumed by the set of dockings given a single metric and plot TPR as a function of FPR in order to obtain an ROC curve. We compute the area under curve (AUC) for ROC curves obtained for each metric to compare different metrics. The AUC is defined as:

$$\mathrm{AUC} = \int_{x=0}^{1} \mathrm{TPR}\left(\frac{1}{\mathrm{FPR}(x)}\right) \mathrm{d}x \tag{3}$$

The TPR and FPR for different thresholds are used to calculate the fraction of models that can be called correct out of all models and the positive predictive value (PPV). The fraction of acceptable and incorrect models are obtained by multiplying the TPR and FPR with the SR. Multiplying the FPR with the SR results in the false discovery rate (FDR) and the PPV can be calculated by dividing the fraction of acceptable models by the sum of the acceptable and incorrect models. The PPV, FDR and SR are defined as:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{FDR} = 1 - \mathrm{PPV} \tag{5}$$

$$\mathrm{SR} = \text{Fraction of predicted models with DockQ} \geq 0.23 \tag{6}$$

pDockQ. As it is not only desirable to know when a model is accurate but also how accurate this model is, we developed a predicted DockQ score, pDockQ. This score is created by fitting a sigmoidal curve (Fig. 2c) using "curve_fit" from SciPy v.1.4.1[56], to the DockQ scores using the average interface plDDT multiplied with the logarithm of the number of interface contacts, with the following sigmoidal equation:

$$\mathrm{pDockQ} = \frac{L}{1 + e^{-k(x - x_0)}} + \mathrm{b} \tag{7}$$

where

$$x = \text{average interface plDDT} \cdot \log(\text{number of interface contacts}) \tag{8}$$

and we obtain $L = 0.724$, $x_0 = 152.611$, $k = 0.052$ and $b = 0.018$.

Analysis of models. To analyse the possibility of determining when AF2 can model a complex correctly, we analyse the structures and the MSAs. We investigated: the number of effective sequences (Neff), the secondary structure in the interface annotated using DSSP[57], the length of the shortest chain, the number of residues in the interface and the number of contacts in the interface.

DSSP was run on the entire complexes, and the resulting annotations were grouped into three categories; helix (3-turn helix (3_{10} helix), 4-turn helix (α helix) and 5-turn helix (π helix)), sheet (extended strand in parallel or antiparallel β-sheet conformation and residues in isolated β-bridges) and loop (residues which are not in any known conformation).

In addition, we assess the PPV of the top N interface DCA signals using the paired MSAs. Here, N is the number of true interface contacts (Cβs from different chains within 8 Å from each other). The PPV is therefore the fraction of the top N DCA signals in the interface that are true contacts. The DCA signals are computed using GaussDCA[58].

$$\text{Interface PPV} = \frac{\text{Number of correct contacts among top N interface DCA signals}}{N} \tag{9}$$

Computational cost. To compare the computation required for each MSA, we compared the time it took to generate MSAs for three protein pairs (PDB: 4G4S_P-O, 5XJL_A-2 and 5XJL_2-M), using either the block diagonalization or AF2 protocol. The tests were performed on a computer using 16 CPU cores from an Intel Xeon E5-2690v4.

Fusing the MSAs took 3 s on average per tested complex. It took 7884 s for generating the AF2 MSAs, the single-chain searches took 338 s on average and the pairing 2 s. The pairing and fusing are thereby negligible compared to searching, resulting in a speedup of 24 times for the hhblits searches. In comparison, folding using the m1-10-1 strategy took 191 s on average for these pairs.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw data used in this study, including multiple sequence alignments and predicted PDB files, are available in the figshare from Science for Life laboratory under accession code 16866202.v1. All other data supporting the findings of this study are available within the article and its supplementary information files. The results used to produce all figures can be found in the supplementary information. Additional information and relevant data will be available from the corresponding author upon reasonable request.

Code availability

All code to run FoldDock and reproduce the analysis here can be obtained here https://gitlab.com/ElofssonLab/FoldDock (commit 2e4c96aa352338976260ece0646ceaaa75392dec) under the Apache License, Version 2.0.

Received: 30 September 2021; Accepted: 11 February 2022;
Published online: 10 March 2022

References

1. Liddington, R. C. Structural Basis of Protein–Protein Interactions. *Protein-Protein Interactions* **261**, 3–14 https://doi.org/10.1385/1-59259-762-9:003 (2004).
2. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* **108**, 1225–1244 (2008).
3. Nooren, I. M. A. NEW EMBO MEMBER'S REVIEW: diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492 (2003).
4. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
5. Zhang, Q. C. et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
6. Marshall, G. R. & Vakser, I. A. Protein-Protein Docking Methods. In *Proteomics and Protein-Protein Interactions* (ed. Waksman, G.) 115–146 (Springer, 2005).
7. Kundrotas, P. J., Zhu, Z., Janin, J. & Vakser, I. A. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl Acad. Sci. USA* **109**, 9438–9441 (2012).
8. Porter, K. A., Desta, I., Kozakov, D. & Vajda, S. What method to use for protein–protein docking? *Curr. Opin. Struct. Biol.* **55**, 1–7 (2019).
9. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443 (2002).
10. Shammas, S. L. et al. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *J. Biol. Chem*, **291**, 6689–6695 (2016).
11. Eginton, C., Naganathan, S. & Beckett, D. Sequence-function relationships in folding upon binding. *Protein Sci.* **24**, 200–211 (2015).
12. Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. Principles of flexible protein-protein docking. *Proteins* **73**, 271–289 (2008).
13. Kurkcuoglu, Z. & Bonvin, A. M. J. J. Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein-protein and protein-DNA systems. *Proteins* **88**, 292–306 (2020).
14. Lensink, M. F. et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* **87**, 1200–1221 (2019).
15. Vreven, T. et al. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* **427**, 3031–3041 (2015).
16. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
18. Kandathil, S. M., Greener, J. G., Lau, A. M. & Jones, D. T. Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterised proteins. *Proc. Natl Acad. Sci. USA* **119**, e2113348119 (2022).

19. Chowdhury, R. et al. Single-sequence protein structure prediction using language models from deep learning. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.02.454840 (2021).
20. Procaccini, A., Lunt, B., Szurmant, H., Hwa, T. & Weigt, M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PLoS ONE* **6**, e19729 (2011).
21. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
22. Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **34**, i802–i810 (2018).
23. Yang, J. et al. Improved protein structure prediction using predicted inter-residue orientations. Preprint at *bioRxiv* https://doi.org/10.1101/846279 (2019).
24. Pozzati, G. et al. Limits and potential of combined folding and docking using PconsDock. *Bioinformatics* **38**, 954–961 (2021).
25. Lamb, J. & Elofsson, A. pyconsFold: a fast and easy tool for modelling and docking using distance predictions. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab353 (2021).
26. Szurmant, H. & Weigt, M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr. Opin. Struct. Biol.* **50**, 26–32 (2018).
27. Green, A. G. et al. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1–12 (2021).
28. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
29. Lensink, M. F. et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins* https://doi.org/10.1002/prot.26222 (2021).
30. Huang, S.-Y. & Zou, X. MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins* **78**, 3096–3103 (2010).
31. Blohm, P. et al. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.* **42**, D396–D400 (2014).
32. Vakser, I. A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* **Suppl 1**, 226–230 (1997).
33. Basu, S. & Wallner, B. DockQ: a quality measure for protein-protein docking models. *PLoS ONE* **11**, e0161879 (2016).
34. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinforma.* **20**, 473 (2019).
35. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
36. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
37. Burke, D. F. et al. Towards a structurally resolved human protein interaction network. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.08.467664 (2021).
38. Kundrotas, P. J. et al. Dockground: a comprehensive data resource for modeling of protein complexes. *Protein Sci.* **27**, 172–181 (2018).
39. Gabler, F. et al. Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinforma.* **72**, e108 (2020).
40. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
41. Rajagopala, S. V. et al. The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* **32**, 285–290 (2014).
42. Kuhlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444 (2014).
43. Orchard, S. et al. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
44. BFD. https://bfd.mmseqs.com/.
45. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
46. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
47. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
48. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
49. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
50. Kosciolek, T. & Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. *Proteins* **84**, **Suppl 1**. 145–151 (2016).
51. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
52. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
53. Singh, A., Dauzhenka, T., Kundrotas, P. J., Sternberg, M. J. E. & Vakser, I. A. Application of docking methodologies to modeled proteins. *Proteins* **88**, 1180–1188 (2020).
54. Anishchenko, I., Kundrotas, P. J. & Vakser, I. A. Contact potential for structure prediction of proteins and protein complexes from Potts model. *Biophys. J.* **115**, 809–821 (2018).
55. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **78**, 3073–3084 (2010).
56. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
57. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
58. Baldassi, C. et al. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS ONE* **9**, e92721 (2014).

Acknowledgements

We thank Petras Kundrotas for supplying the new heterodimeric proteins without templates in the PDB. We also thank Liming Qiu and Xiaoqin Zou for their help with running their docking program MDockPP in a timely manner. Financial support: Swedish Research Council for Natural Science, grant No. VR-2016-06301 and Swedish E-science Research Center. Computational resources: Swedish National Infrastructure for Computing, grants: SNIC 2021/5-297, SNIC 2021/6-197 and Berzelius-2021-29. All financial support and computational resources were received by A.E.

Author contributions

P.B. and G.P. performed the studies; all authors contributed to the analysis. P.B. wrote the first draft of the manuscript; all authors contributed to the final version. A.E. obtained funding.

Funding

Open access funding provided by Stockholm University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28865-w.

Correspondence and requests for materials should be addressed to Patrick Bryant or Arne Elofsson.

Peer review information *Nature Communications* thanks Rodrigo Honorato and Shoshana Wodak for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022, corrected publication 2022

nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-022-00910-8

Towards a structurally resolved human protein interaction network

Received: 11 February 2022

Accepted: 14 December 2022

Published online: 23 January 2023

Check for updates

David F. Burke[1,9], Patrick Bryant[2,3,9], Inigo Barrio-Hernandez[1,9], Danish Memon[1,9], Gabriele Pozzati[2,3,9], Aditi Shenoy[2,3], Wensi Zhu[2,3], Alistair S. Dunham[1], Pascal Albanese[4,5], Andrew Keller[6], Richard A. Scheltema[4,5], James E. Bruce[6], Alexander Leitner[7], Petras Kundrotas[2,3,8] ✉, Pedro Beltrao[1,7] ✉ & Arne Elofsson[2,3] ✉

Cellular functions are governed by molecular machines that assemble through protein-protein interactions. Their atomic details are critical to studying their molecular mechanisms. However, fewer than 5% of hundreds of thousands of human protein interactions have been structurally characterized. Here we test the potential and limitations of recent progress in deep-learning methods using AlphaFold2 to predict structures for 65,484 human protein interactions. We show that experiments can orthogonally confirm higher-confidence models. We identify 3,137 high-confidence models, of which 1,371 have no homology to a known structure. We identify interface residues harboring disease mutations, suggesting potential mechanisms for pathogenic variants. Groups of interface phosphorylation sites show patterns of co-regulation across conditions, suggestive of coordinated tuning of multiple protein interactions as signaling responses. Finally, we provide examples of how the predicted binary complexes can be used to build larger assemblies helping to expand our understanding of human cell biology.

Proteins are key cellular effectors determining most cellular processes. These rarely act in isolation, but instead, the coordination of the diversity of processes arises from the interaction among multiple proteins and other biomolecules. The characterization of protein-protein interactions (PPIs) is crucial for understanding which groups of proteins form functional units and underlies the study of the biology of the cell. Diverse experimental and computational approaches have been developed to determine the PPI network of the cell (that is, the interactome), with hundreds of thousands of human protein interactions determined to date[1–3]. Protein interactions vary from transient interactions that regulate an enzyme to permanent interactions in molecular machines.

The structural characterization of the human interactome has lagged behind, with experimental and homology models currently covering an estimated 15 protein interactions[4,5]. The structural characterization of protein complexes is a critical step in understanding the mechanisms of protein function, and in studying the impact of mutations[4,6–8] and the regulation of cellular processes via the post-translational tuning of binding affinities[9–12].

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [2]Science for Life Laboratory, Stockholm University, Solna, Sweden. [3]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden. [4]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands. [5]Netherlands Proteomics Center, Utrecht, The Netherlands. [6]Department of Genome Sciences, University of Washington Seattle, Seattle, WA, USA. [7]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [8]Center for Computational Biology, The University of Kansas, Lawrence, KS, USA. [9]These authors contributed equally: David F. Burke, Patrick Bryant, Inigo Barrio-Hernandez, Danish Memon and Gabriele Pozzati. ✉e-mail: pkundro@ku.edu; pbeltrao@ebi.ac.uk; arne@bioinfo.se

Computational approaches for predicting the structures of interacting protein pairs are primarily based on identifying structural similarity for pairs of proteins against experimentally determined protein complexes[4,6,13,14]. The Interactome3D (refs. [4,14]) repository currently lists 7,625 predicted models based on homology of domains, a number similar to the 8,359 pairs listed having an experimentally determined model. In addition, co-evolution-based information has been used to predict protein interactions and to guide structural docking for bacterial proteins[15]. Recently, neural network-based approaches have demonstrated the ability to accurately predict the structures of individual proteins[16,17] and protein complexes[16,18–21]. These approaches can correctly predict the structures of up to 60% of dimers[18], and have been used to predict structures of 1,506 *Saccharomyces cerevisiae* protein interactions[22]. However, the application of these neural network models for the large-scale prediction of human complex structures has not been tested yet.

Here, we assess the possibilities and limitations of applying AlphaFold2 to modeling human protein interactions on a large scale. We predicted the complex structures for two sets of human interactions obtained using different experimental methods, comprising 65,484 unique human interactions. We show that it is possible to rank the models according to confidence, with 3,137 predicted structures ranked as highly confident. Further, we show that the higher-confidence predictions are enriched among those supported by a combination of experimental methods. We showcase the value of a structurally resolved interactome by studying disease mutations and phosphorylation of interface residues. Finally, we provide some indication that binary complexes can be used to build higher-order assemblies.

Structure prediction of human protein interactions

We selected experimentally identified human protein interactions from the Human Reference Interactome (HuRI)[2] and the Human Protein Complex Map (hu.MAP v.2.0)[3]. HuRI comprises protein interactions determined by yeast two-hybrid (Y2H) screening[2] from which we modeled 55,586 pairs. From hu.MAP we selected 10,207 high-quality PPIs[3]. While HuRI is more likely to be enriched for direct protein interactions, including transient partners, the hu.MAP set is more likely to reflect stable protein interactions, including members of the same complex that may not be interacting directly. The overlap between the two datasets is small (309 pairs), and a comparison with two large-scale compendiums of structural models[4] indicates that 62,019 of the combined pairs do not have experimental models nor can they be modeled easily by homology, suggesting a large potential gain in structural knowledge.

We predicted the structure of 65,484 nonredundant pairs using the FoldDock pipeline[18], based on AlphaFold2 (ref. [17]). As in the FoldDock pipeline, we combined size and the predicted local Distance Difference Test (plDDT) scores of the interface into a single score to predict the DockQ score of a complex, dubbed pDockQ (Methods), which can rank models by confidence. We tested pDockQ score by comparing the predicted models with 1,465 experimental models, of which 742 (50%) were correct (DockQ > 0.23). For predictions with pDockQ > 0.23, 70% (671 of 955) are well modeled, and for pDockQ > 0.5, 80% (521 of 651).

We show in Fig. 1a the distribution of pDockQ for the predicted and random protein interactions, and provide data for all models in Supplementary Table 1. The pDockQ of known interacting proteins tends to be higher than for the random set, with the predictions for hu.MAP showing on average higher confidence than the HuRI set. Additionally, when selecting hu.MAP interactions also supported by Y2H or crosslink data (crosslinking) results in even higher-confidence values (Fig. 1a). This suggests that high-confidence models are enriched for protein interactions supported by the two types of methods associated with high affinity and direct interactions. We identified 3,137 structures (Fig. 1b) as high-confidence models (pDockQ > 0.5). The number of structures increased to 10,061 if a cut-off of 0.23 was used. Only 0.3% of the random set of models would be considered confident predictions at this cut-off. In Fig. 1c we show examples of predicted structures aligned to experimental or homology models, showing how the predictions and the confidence score relate to the observed alignments. For the majority of these cases, even with lower-confidence values, the interaction interface is generally in good agreement, except for the interaction between subunits of the proteasome 26S complex, ATPase domain 2 (PSMC2) and non-ATPase domain 11 (PSMD11). It can be noted that several of the models in Fig. 1c are parts of large complexes: PRDX2–PRDX3: members of the peroxiredoxin family of antioxidant enzymes; RFC2–RFC5: subunits of heteropentameric Replication factor C (RF-C); YWHAB–YWHAG: parts of the 14-3-3 family of proteins tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation proteins beta (YWHAB) and gamma (YWHAG); and RPL9–RPL18A: ribosomal proteins L9 (RPL9) and L18a (RPL18A). This shows that AlphaFold2 can predict the structures of directly interacting protein pairs present in large complexes.

Features impacting prediction confidence

As shown in Fig. 1a, protein pairs present in the Protein Data Bank (PDB) are enriched in high-scoring models compared with pairs in HuRI and Hu.MAP. There could exist several possible explanations for this, such as the inability of AlphaFold2 to identify transient or indirect interactions. Nevertheless, it is also possible that the two high-throughput datasets contain noninteracting pairs. Therefore, to understand this difference better, we first studied an additional dataset created from large (>10 chains) heteromeric protein complexes.

The set of large complexes consists of 12 large heteromeric protein complexes, and all (nonidentical) pairs of protein chains in each complex were docked with each other. These pairs can be divided into the ones with direct interaction and those that do not interact directly. Here, we used a definition of more than 20 contacts of less than 8 Å between Calphas to exclude small interaction interfaces. When a complex contained multiple copies of identical chains, all interactions were included to allow for alternative interactions between the chains. The difference in pDockQ scores between the direct and indirect interacting pairs is striking, where only 6% of the indirect pairs have a pDockQ score > 0.5 compared with 38% of the directly interacting pairs (Fig. 2a). This shows that directly interacting pairs often can be predicted even when they are part of large complexes, in contrast to indirectly interacting pairs.

hu.MAP has many more high-confidence predictions than HuRI, which is based on Y2H experiments. To further understand this difference, we first analyzed a subset of all protein pairs from the CORUM[23] database, the best manually curated database of mammalian protein complexes, and predicted the interaction of all pairs in the same complex. The average pDockQ score of CORUM is slightly higher than for hu.MAP, but the number of high-quality predictions is similar (16% versus 19%), indicating that the different databases of protein complexes have a similar fraction of high-confidence predictions and that HuRI is the outlier (Fig. 2b).

It is unlikely that the Y2H in HuRI data should contain a large set of indirect interactions, as only two human proteins are expressed in the same cell. Therefore, there must be another reason for the few high-confidence predictions. We examined the properties of the pairs present in the two datasets. Here, it can be seen that HuRI proteins differ from the hu.MAP (and other datasets) in two ways. HuRI protein pairs contain more intrinsic disorder (Fig. 2c) and have fewer efficient sequences (meff) in their multiple sequence alignments (MSAs) (Fig. 2d). In these figures it can also be seen that the pDockQ values tend to increase with less disorder and more sequences in the alignments, although it is clearly not an absolute relationship. Further, protein pairs in HuRI are less likely to be found in the same subcellular compartment (Fig. 2e), and have similar coexpression profiles (Fig. 2f). Considering all this, it is likely that many protein interactions in HuRI are transient and that AlphaFold2 cannot reliably predict such interactions.

https://doi.org/10.1038/s41594-022-00910-8

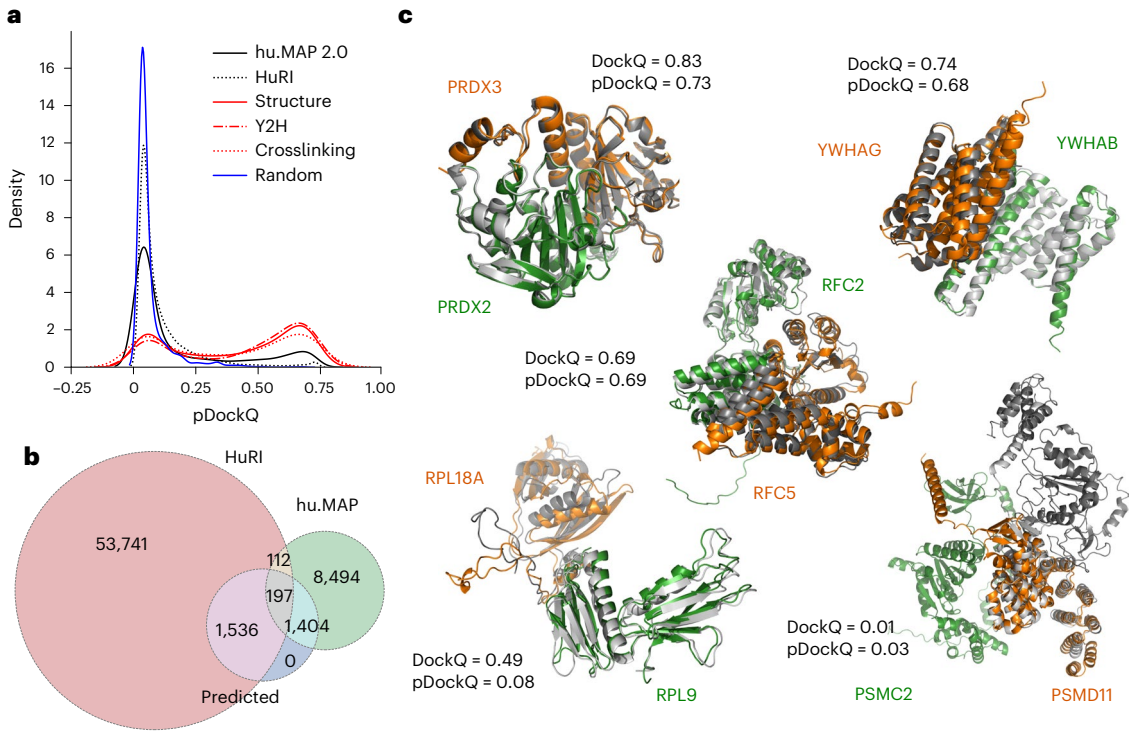


Fig. 1 | Application of AlphaFold2 complex predictions to a large dataset of human PPIs. a, Distribution of model confidence score (pDockQ) for predicted structures from two large human protein interaction datasets (hu.MAP and HuRI), compared with confidence metrics from 2,000 random pairs of proteins. The hu.MAP dataset was further subsetted to those that have support from Y2H ('Y2H') or crosslink data ('Crosslinking'), or correspond to pairs with available experimental or homology modeling information ('Structure'). **b**, Number of protein interactions with models built from both datasets and those that we consider being of high confidence ('Predicted'), corresponding to those with pDockQ > 0.5. **c**, Examples of predicted models (orange and green) overlapped with the corresponding experimental models (gray) and the observed (DockQ) or predicted (pDockQ) quality of the models.

Crosslinking support for predicted complex structures

Chemical crosslinking followed by mass spectrometry is an approach which can be used to identify reactive residues (usually lysines) that are in proximity, as constrained by the geometry of the crosslink agent used. The identification of such residues across a pair of proteins can help define the likely protein interface. To determine if the predicted complex structures agree with such orthogonal spatial constraints, we obtained a compilation of crosslinks for pairs of residues across 528 protein pairs with predicted models (Fig. 3a, Supplementary Table 1 and Methods). In total, 51% of the models had one or more crosslinks at a distance below the expected maximal distance possible (Fig. 3a). Restricting the predicted models to higher confidence by the pDockQ score increased the fraction of complexes with acceptable crosslinks, reaching 75% for pDockQ scores greater than 0.5 (Fig. 3a). This result is in line with the benchmark results above.

In total, we have identified 479 crosslinks providing supporting evidence for 171 predicted complex structures with pDockQ > 0.5. Of these, 41 correspond to complex structures with no experimental structure or homology models, from which we selected some to illustrate (Fig. 3b–e). Figure 3b shows the AlphaFold2 (AF2) model for the full length of the ERLIN1/ERLIN2 complex, which mediates the endoplasmic reticulum-associated degradation (ERAD) of inositol 1,4,5-trisphosphate receptors (IP3Rs). AlphaFold2 predicts a globular domain (1–190) followed by an extended helical region with a kink around amino acid position 280. Unlike the model in Interactome3D, the paralogous proteins are stacked side-by-side with the hydrophobic face of the helices buried and the hydrophilic face (mainly Lys) exposed to solvent. A crosslink between the C-terminal residues K275 (ERLIN1) and K287 is predicted to bridge a distance of 18 Å, supporting the predicted model. In Fig. 3c we show the model for proteins IMMT and CHCHD3, components of the mitochondrial inner membrane MICOS complex. AlphaFold2 predicts a globular helical domain at the C-terminal end of IMMT (550–750) to interact with the C-terminal end of CHCHD3 (150–225). This is supported by data of three crosslinks: between K173 (CHCD3) and K565 (IMMT), and K203 (CHCD3) to both K714 and K726 of IMMT. Figure 3d shows the complex of transfer RNA-guanine-*N*(7)-methyltransferase (METTL) with its noncatalytic subunit (WDR4). The structure of WDR4 has not yet been solved experimentally but contains WD40 repeats, which are expected to form a β-propeller domain, as predicted here. The METTL domain is predicted to interact with the side of the WDR40, away from the ligand-binding pore. This orientation is supported by a crosslink between K122 (WDR4) and K143 (METTL) (18 Å). Finally, in Fig. 3e we show the predicted complex structure for the heterogeneous nuclear ribonucleoprotein C (HNRNPC) and the RNA-binding protein, RALY. Two regions in both proteins are predicted with high confidence (plDDT > 70), with the lower-confidence regions not shown. The N-terminal domain in HNRNPC (16–85) is predicted to interact with the N-terminal domain of RALY (1–100). A long helix in HNRNPC (185–233) is predicted to interact with a helix in RALY (169–228). This interhelix interface is supported by crosslinking data for three pairs of lysines at either end of the helices (189 → 222; 229 → 179; and 232 → 183).

Disease-associated missense mutations at interfaces

Missense mutations associated with human diseases can alter protein function via diverse mechanisms, including disrupting protein stability, allosterically modulating enzyme activity and altering PPIs. Structural models can allow the rationalization of possible mechanisms

https://doi.org/10.1038/s41594-022-00910-8

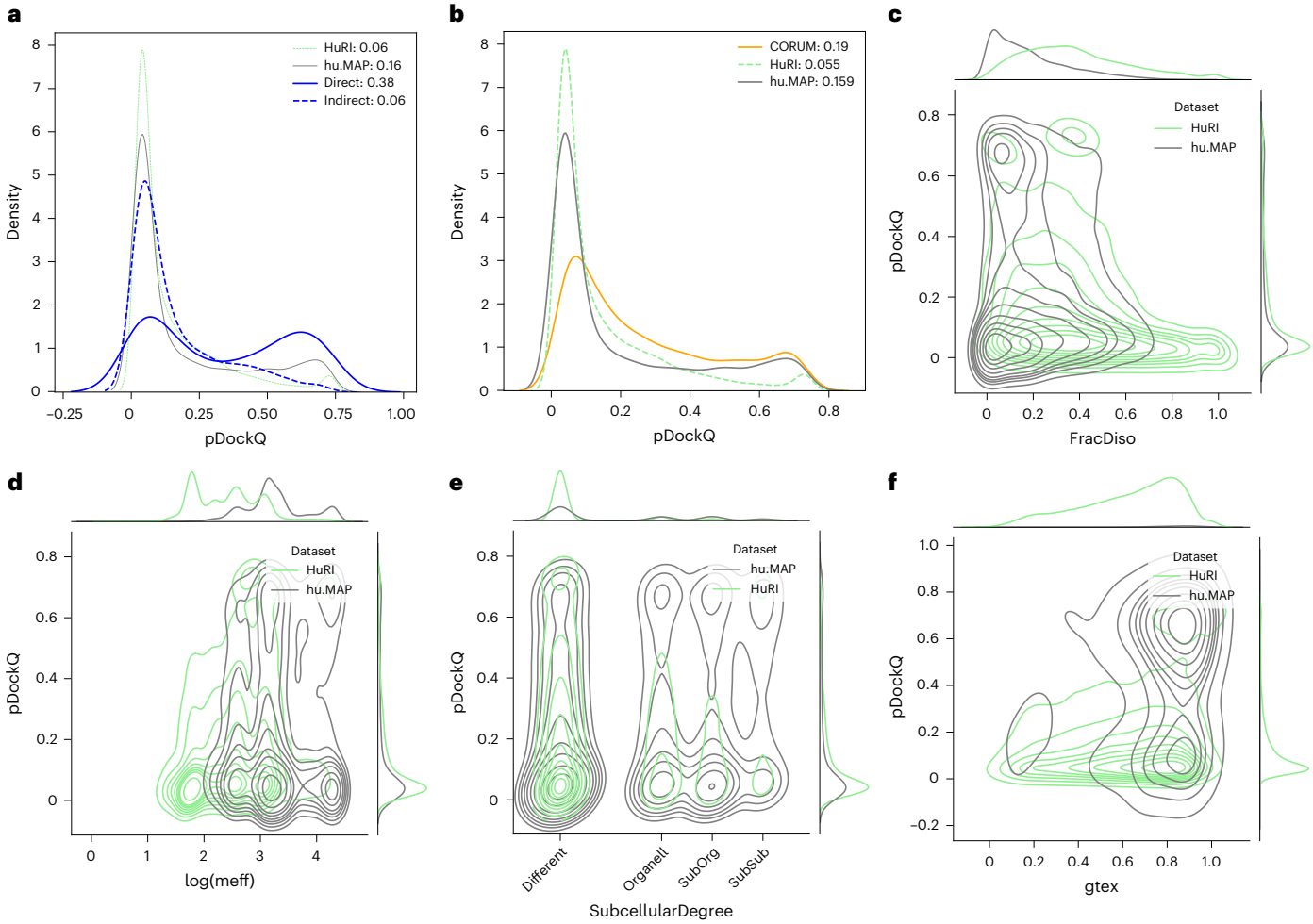


Fig. 2 | Protein and interaction features impacting on prediction confidence: analysis of different datasets. In all subfigures, proteins from HuRI in green, hu.MAP gray, CORUM orange and from large PDB complexes blue. **a**, pDockQ values of directly and indirectly interacting proteins from the same complex (blue); for comparison, HuRI and hu.MAP data are shown with thin lines. **b**, pDockQ values of CORUM (orange), HuRI (green) and hu.MAP (gray) datasets. **c**, Fraction of residues predicted to be disordered (pLDDT < 0.5) shows that protein pairs in HuRI are enriched in disorder. **d**, Proteins in HuRI have fewer sequences in the paired MSAs as measured by the mean number of efficient sequences in the MSA (meff). **e**, Proteins that share subcellular localization (solid lines) are enriched in high pDockQ scores in all three datasets. **f**, Only protein pairs in hu.MAP are coexpressed according to STRING, using similarity in Genotype-Tissue Expression (gtex), and coexpressed pairs are enriched in pairs with high pDockQ scores.

of interface disease mutations. To determine the usefulness of the predicted structures, we compiled a set of mutations located at interface residues that were previously experimentally tested for the impact on the corresponding interaction[24]. We then performed in silico predictions of changes in binding affinity upon mutations using FoldX[25] and observed that mutations known to disrupt the interactions are predicted to have a strong destabilization of binding compared with mutations known not to have an effect (Fig. 4a and Supplementary Table 2). Very high confidence (plDDT > 90) of the mutated residues led to more substantial discrimination between mutations known and not known to disrupt the complex formation (Fig. 4a), indicating that only very accurate models are useful when using the FoldX forcefield for estimating the impact of binding affinity of mutations.

Next, we mapped human disease (from ClinVar) and cancer mutations (from The Cancer Genome Atlas) to the interface residues defined by the set of high-confidence protein complex predictions (pDockQ > 0.5) (Supplementary Table 1). The hu.MAP and HuRI confident predictions identified 280 interfaces carrying pathogenic mutations and 602 interfaces corresponding to the top 25% of recurrently mutated interfaces in cancer, defined as the highest number of mutations per interface position (Fig. 4b and Methods). We find a strong enrichment in pathogenic versus benign mutations at interface residues relative to the rest of the protein (2.3-fold enrichment, P value 2.7×10^{-31}).

We illustrate in Fig. 4c examples of protein network clusters with interface disease mutations across a range of biological functions. For example, interface mutations in chromatin remodeling, including members of SWI/SNF complex (SMARCD1, SMARCD2, SMARCD3), and several transcription factors related to development (for example, TCF3, TCF4, LMO1 and LMO2).

We selected examples of interfaces with disease mutations and no previous experimental data or homology to available models (Fig. 4d–g). Figure 4d shows the interface of WDR4-METTL1, which has supporting crosslink information described above. WDR4 has two annotated pathogenic variants at this interface, linked with Galloway-Mowat Syndrome 6, with the highlighted R170 participating in interactions with a negatively charged residue of METTL1. Figure 4e shows an example of an interface with 32 recorded interface mutations in cancer for both proteins, including the highlighted arginines in LDOC1, which form electrostatic interactions with the opposite chain. TWIST1 has several annotated pathogenic mutations, including L149R and L159H, which

https://doi.org/10.1038/s41594-022-00910-8

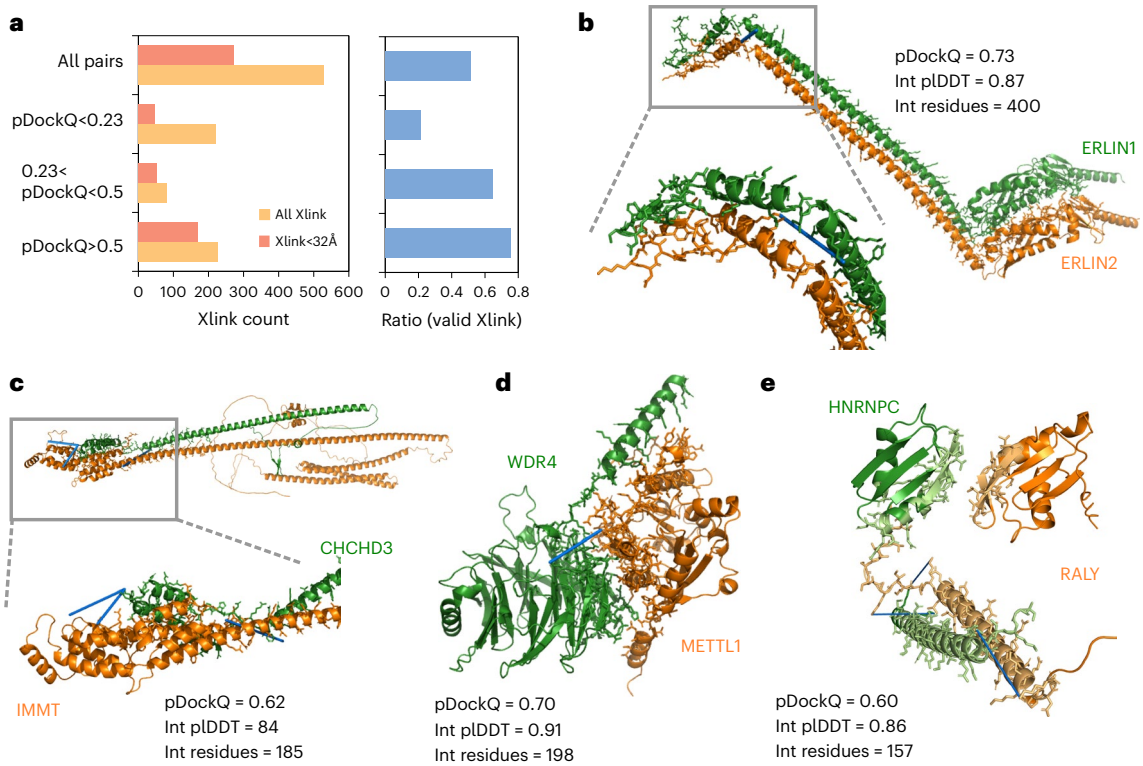


Fig. 3 | Crosslink support for predicted complex models. a, The numbers and ratios of predicted structures having crosslink information for pairs of residues that bridge the two proteins in the predicted structure, broken down by the crosslinks that satisfy their expected maximal distance and by the predicted quality of the model (pDockQ). **b**–**e**, Examples of predicted structures of high confidence, with no previous structural information and supported by at least one crosslink (indicated with blue line): ERLIN1/ERLIN2 complex (**b**), IMMT and CHCHD3, components of the mitochondrial inner membrane MICOS complex (**c**), the complex of transfer RNA-guanine-N(7)-methyltransferase (METTL) with its noncatalytic subunit (WDR4) (**d**) and the heterogeneous nuclear ribonucleoprotein C (HNRNPC) and the RNA-binding protein, RALY (**e**).

are at residues buried in the interface (Fig. 4f). In particular, the L149R mutation, associated with Saethre–Chotzen syndrome, would strongly disrupt packing. The R118G mutation would disrupt the interaction with residue F22 mainchain O in TCF4. In RAD51D we found the mutation R266C (Breast-ovarian cancer, familial), which interacts across the interface with XRCC2 (Fig. 4g) and paralogous genes involved in the repair of DNA double-strand breaks by homologous recombination. Interestingly, we also found mutations at R239, to Trp/Gln/Gly, associated with Breast-ovarian cancer which interacts with Tyr119 in XRCC2, which itself is also annotated as having mutations linked to hereditary cancer-predisposing syndrome.

Phospho-regulation of protein complex interfaces

Protein phosphorylation can regulate protein interactions by modulating the binding affinity via the change in size and charge of the modified residue. Over 100,000 experimental human phosphorylation sites have been determined to date[26,27], but only 5–10% of these have a known function[28]. Mapping phosphorylation site positions to protein interfaces can generate mechanistic hypotheses for their functional roles in controlling protein interactions. We used a recent characterization of the human phosphoproteome[26] to identify 4,145 unique phosphosites at interface residues among the highly confident models. The average functional importance, defined by the functional score described earlier[26], was generally higher than random for phosphorylation sites at interfaces (Fig. 5a), and we found some enrichment for targets of multiple kinases, including tyrosine kinases (ERBB2, AXL, ABL2, FER) (Fig. 5b). This suggests that some interfaces may be under coordinated regulation by specific kinases and conditions.

To identify potentially co-regulated interfaces, we collected measurements of changes in phosphorylation levels across a large panel of over 200 conditions[29]. We retained 260 phosphosites that had a significant regulation in three conditions and then computed all-by-all pairwise correlations in phosphosite fold changes across conditions (Supplementary Table 1). We clustered these phosphosites by their profile of correlations (Fig. 5c), identifying 16 groups of co-regulated interface phosphorylation sites (Fig. 5c and Supplementary Table 3). For each group of phosphosites, we identified the conditions where these have the strongest up- or down-regulation (Supplementary Fig. 1) and plotted a subset of conditions in Fig. 5d. We also performed a gene ontology enrichment analysis for each group of co-regulated phosphosites, including both proteins of the modified interfaces, to search for common biological functions (Fig. 5e and Supplementary Table 4). Here, one-sided hyper-geometric tests were used for statistical analysis. For example, we observed a cluster of interface phosphosites in proteins related to intermediate filaments (cluster 7) which show strong regulation patterns along the cell cycle, downregulated in S-phase and up-regulated in G1 and mitosis. Phosphosites in cluster 1 (cell cycle G1-S phase transition) show the opposite trends, with up-regulation in late S-phase and down-regulation in G1 and mitosis. Some clusters show regulation under specific kinase inhibition, which may provide novel hypotheses for kinase regulation of specific processes. For example, phosphosites in cluster 9 (regulation of chromosome assembly) tend to be up-regulated after inhibition of ROCK and up-regulated after inhibition of mTOR.

While not all phosphosites at interfaces are likely to regulate the binding affinity, this analysis provides hypotheses for the potentially coordinated regulation of multiple proteins by tuning of their interactions after specific perturbations.

https://doi.org/10.1038/s41594-022-00910-8

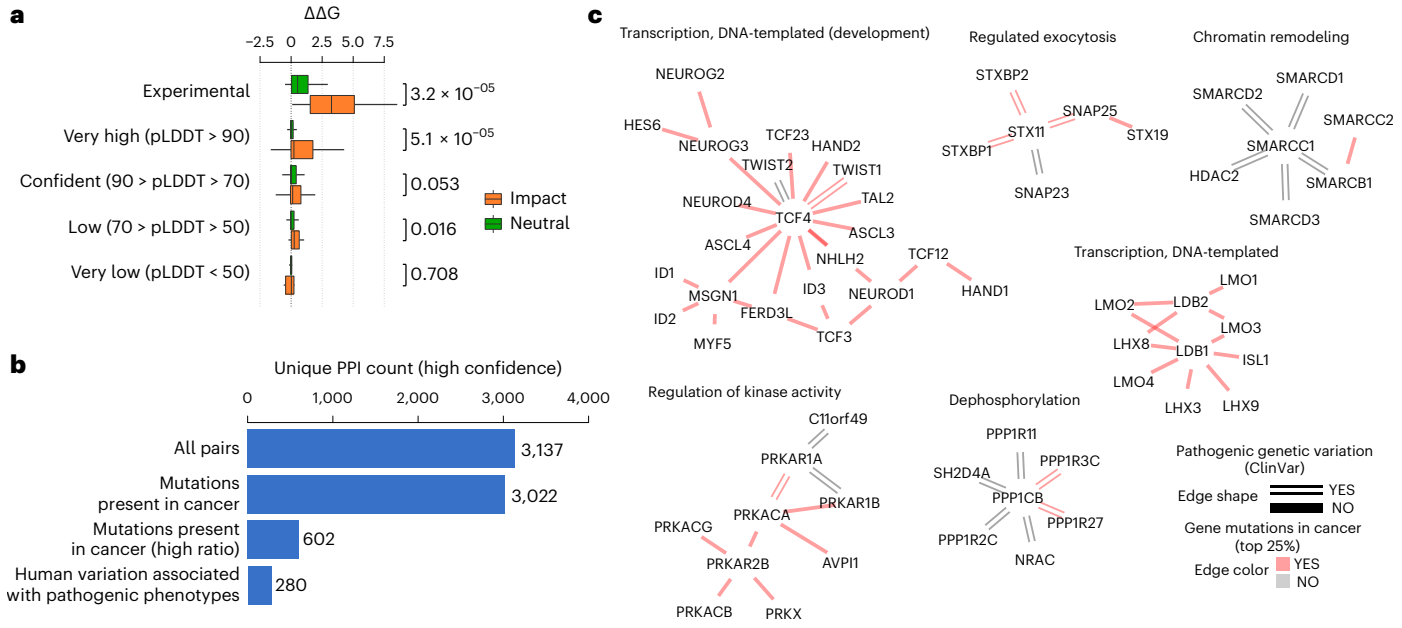


Fig. 4 | Disease mutations at protein complex interface residues. a, Boxplot showing the distribution of changes in predicted binding affinity (ΔΔG) for mutations known to have an impact (orange) versus the ones with neutral effect (green). The boxes represent the first and third quartiles. The upper whisker extends from the third quartile to the largest value no further than 1.5 × inter-quartile range (IQR). The lower whisker extends from the first quartile to the smallest value at most 1.5 × IQR. The statistical significance of the differences was tested using two-sided Wilcoxon rank sum test with continuity correction. The total number of data samples was 1,320 (900 neutral, 420 impact) (Supplementary Table 2). The min, mean and max values for the subsets are as follows: Experimental Neutral (−0.5, 1.0, 5.1); Experimental Impact (0.1, 3.4, 8.6); AF plddt > 90 Neutral (−0.9, 0.3, 4.7); AF plddt > 90 Impact (−2.7, 1.0, 12.5,); AF 70 < plddt > 90 Neutral (−2.0, 0.3, 9.0); AF 70 < plddt > 90 Impact (−4.0, 0.5, 7.0); AF 50 < plddt > 70 Neutral (−1.2, 0.1, 2.7); AF 50 < plddt > 70 Impact (−0.2, 0.3, 1.7); AF plddt < 50 Neutral (−2.6, 0.1, 2.0); AF plddt < 50 Impact (−2.6, −0.3, 0.3). **b**, Unique PPI pairs for high-confidence models (pDockQ > 0.5) in total, with mutations in cancer, mapped to the interface (all and top 25% ratios) and with pathogenic or likely pathogenic clinical variants mapped to the interface. **c**, Modules related to relevant biological processes. The color of the edge represents the presence of cancer mutations in the interface (top 25% ratio, color red), and the shape the presence of pathogenic clinical variants (double line). **d**–**g**, Selected relevant structures with no previous structural knowledge showing clinical variants or mutations in cancer mapped to the interface (mutated residues in red). The interface of WDR4-METTL1 (**d**), the interface of LDOC1 (**e**), the interface of TWIST1 (**f**) and the interface of RAD51D (**g**).

Higher-order assemblies from binary protein interactions

Proteins interact with multiple partners either simultaneously, as part of larger protein complexes, or separated in time and space. This is also reflected in our structurally characterized network, where proteins can be found in groups, as illustrated in a global network view of the protein interactions with confident models (Fig. 6, Supplementary Fig. 2 and Supplementary Data 1). One key benefit of structurally characterizing an interaction network is the identification of shared interfaces for multiple interactors. As an example, we highlight GDI1 (RabGDP dissociation inhibitor alpha) which interacts with multiple Rab proteins, regulating their activity by inhibiting the dissociation of GDP. The predicted complex structures for these interactions show how these share the same interface and therefore cannot co-occur. Other clusters in the network suggest that the proteins form larger protein complex assemblies with many-to-many interactions. As the use of AlphaFold2 for predicting larger complex assemblies can be limited by computational requirements, we tested whether the structures for pairs of proteins could be iteratively structurally aligned. We tested this procedure on a small set of complexes covered in this network, with known structures and the number of subunits ranging from five (RFC complex, TFIIH core complex) to 14 (20S proteasome). We then aligned an experimentally determined structure with the predicted models (Fig. 6; gray, experimental model). These examples showcase the potential and also limitations of this procedure.

The TFIIH core complex is composed of five subunits with 1-to-1 stoichiometry. All subunits can be modeled, with the final complex generally agreeing (Fig. 6) with a cryoEM structure for these subunits (PDB:6NMI). The most significant difference to the cryoEM model is the relative positioning of the ERCC3 subunit. The exact final model

https://doi.org/10.1038/s41594-022-00910-8

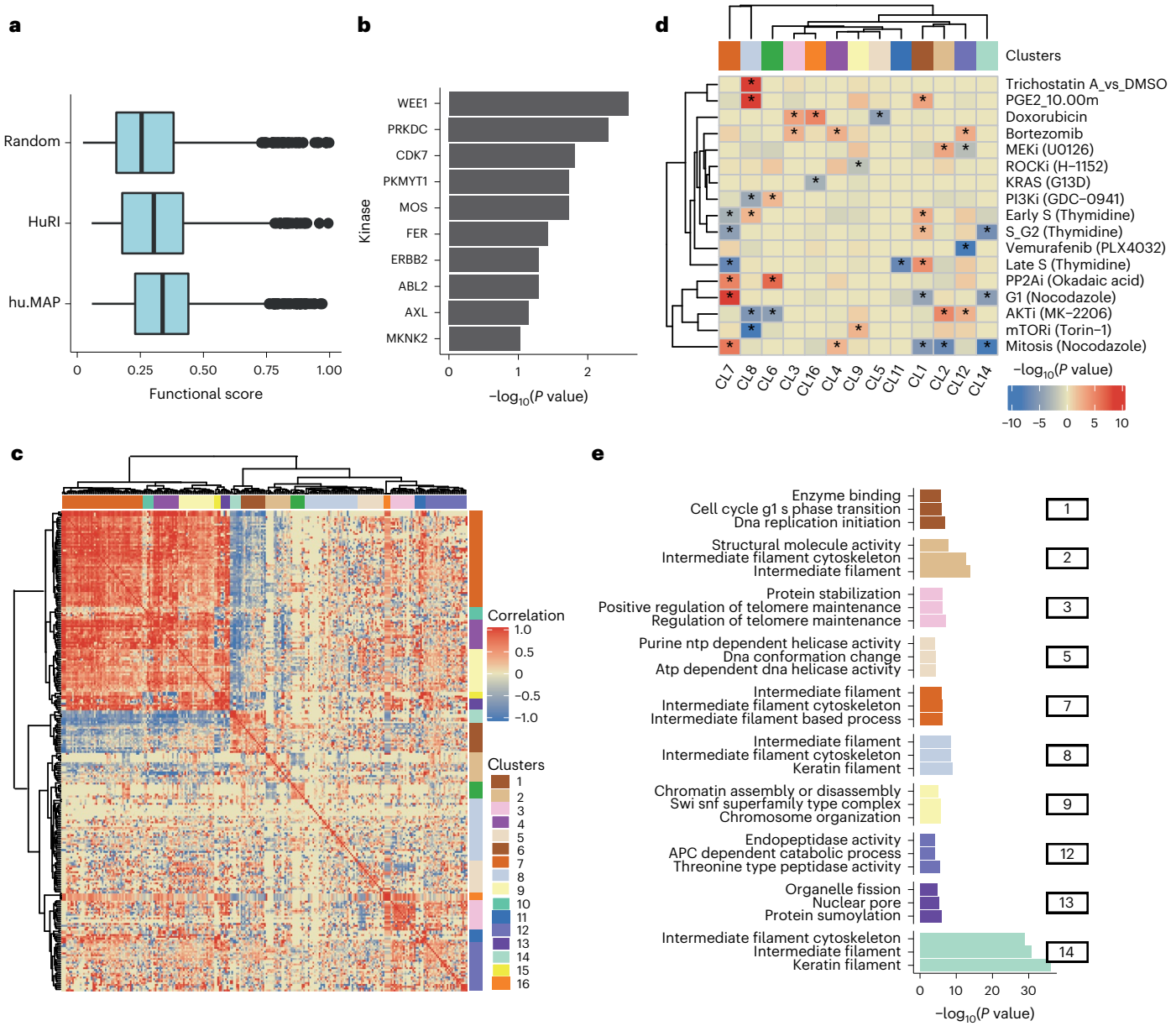


Fig. 5 | Co-regulation of phosphorylation sites at interface residues. **a**, Distribution of phosphosite functional scores for phosphosites at interface residues and random phosphosites. The min, mean and max values were as follows: Random = 0.02, 0.26, 0.98; HuRI = 0.06, 0.37, 0.99; hu.MAP = 0.06, 0.33, 0.99. The boxes represent the first and third quartiles. The upper whisker extends from the third quartile to the largest value no further than 1.5 × IQR. The lower whisker extends from the first quartile to the smallest value at most 1.5 × IQR. **b**, Enrichment of kinase substrates among phosphosites at interface residues. The *P* value was derived from an over-representation analysis using a one-sided hyper-geometric test ($N = 7{,}150$). **c**, Hierarchical clustering of the pairwise correlation values for changes in phosphosite levels across conditions. Groups of phosphosites showing high correlation values were defined as clusters (1 to 16), as indicated in colors along the outside of the clustergram. **d**, Degree of regulation of phosphosites from each cluster in a select panel of conditions, defined by a one-sided *Z*-test comparing the fold change of the phosphosites in a cluster compared with the entire distribution of fold changes in that condition. The result is summarized as the −log(*P* value), and signed as positive if the median value is above the background or negative otherwise. **e**, Gene ontology enrichment analysis for the proteins with phosphosites annotated to select clusters.

obtained can vary depending on the aligned pairs, with multiple possible final conformations (Supplementary Fig. 3). Figure 6 illustrates the conformation that best matches the cryoEM model in PDB:6NMI. For example, for the TFIIH core complex, there is a predicted model where the complex adopts a more open conformation (as seen in PDB:5OQJ) and alternative predicted placements of the GTF2H1 subunit.

The RFC complex is also composed of five subunits with 1-to-1 stoichiometry. One iterative alignment of pairwise protein interactions builds a model that includes all five subunits organized similarly to that observed in the PDB:6VVO cryoEM structure (Fig. 6). In this predicted model, the subunits RFC2/5/4/3 match the experimentally observed model well, but there are apparent deviations introduced by compounding errors in alignment by this iterative process. Individual subunits in the cryoEM structure can be aligned to each of the model subunits well, but then the alignment of the rest of the model is progressively worse the further away the subunits are positioned from the aligned subunit. The RFC1 subunit is individually not well predicted. Further, the RFC3-RFC5 interaction pair is predicted with high confidence, while, in fact, these do not share a direct contact in the experimental structure. AlphaFold2 places RFC3 at the RFC5-RFC4 interface, likely due to the structural similarity between RFC3 and RFC4.

https://doi.org/10.1038/s41594-022-00910-8

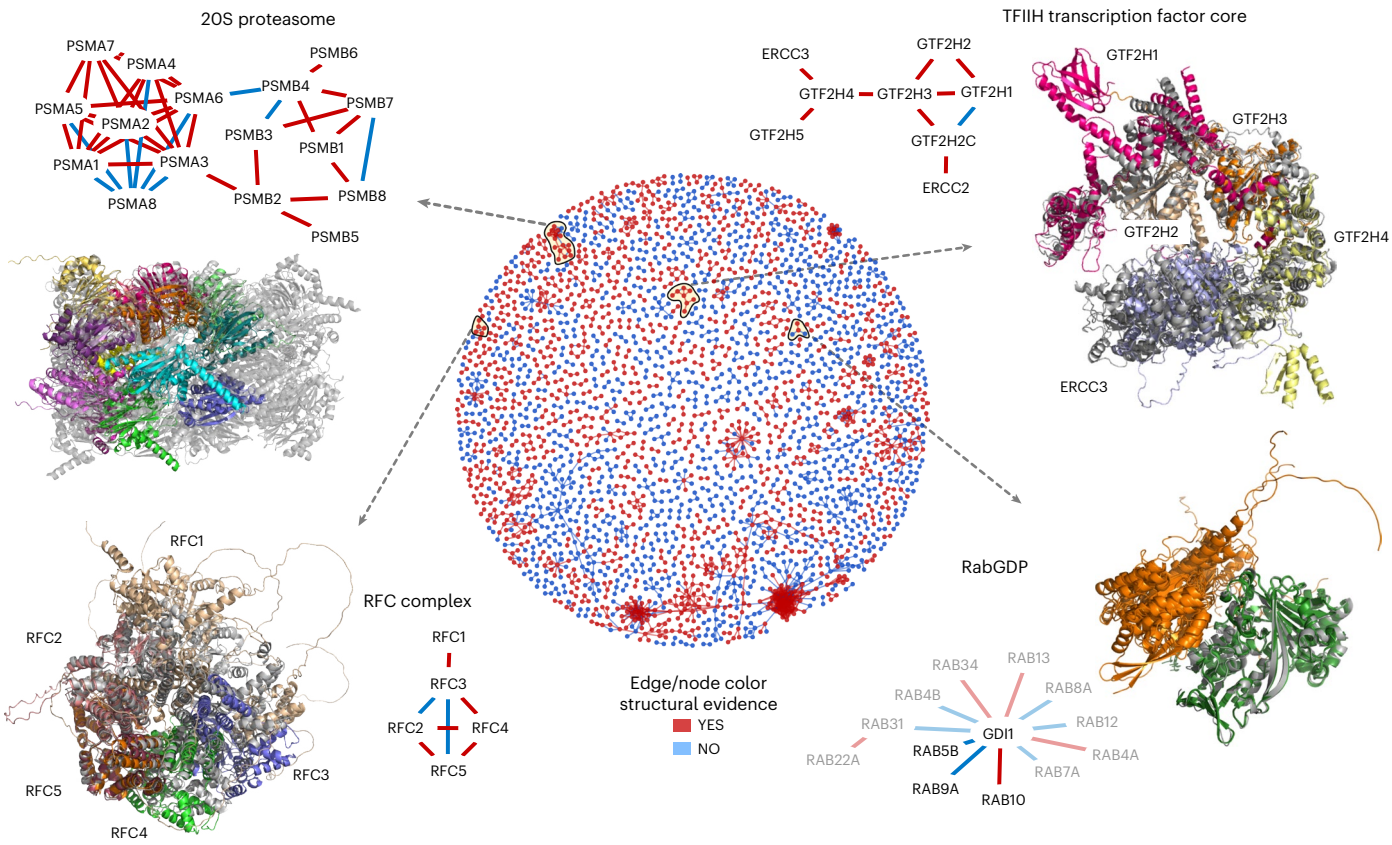


Fig. 6 | Protein complex predictions for higher-order assemblies. The middle circle is a network view of all PPIs predicted with high confidence (pDockQ > 0.5). The edges and nodes are colored in red if there is a previous experimental or homology model for the interaction and blue if such information is unavailable. We selected four examples of recapitulated complexes (yellow circles and black arrows) plotted in further detail. In these small networks, only the edges are colored based on structural evidence. In the case of RabGDP, the faded nodes and edges represent predictions with slightly lower confidence (pDockQ > 0.3).

Encouraged by the examples tested, we defined an automatic procedure to generate larger models by iterative alignment of pairs (Methods). We start building all possible dimers in a complex, then sort them by pDockQ, and start building from the first ranked dimers. Next, we add the highest-ranked dimer, which shares one subunit with the complex if it does not overlap; this is repeated for all dimers until the complex is complete or no additional proteins can be added. We tested this on the 20S proteasome, a particularly challenging example, with stoichiometries different from 1-to-1 and homologous subunits. This automatic procedure could build a model containing all 14 subunits (half of the proteasome), which are mostly placed in agreement within the experimental model (Fig. 6). However, the exact order of the chains is incorrect, that is, at each location an incorrect protein is placed, highlighting that AF2 cannot distinguish which two proteins interact from a set of homologous proteins.

Two additional proteins where we could build a good model are Heterodisulfide reductase from *Methanothermococcus thermolithotrophicus* (PDB:5ODC) and the eukaryotic translation initiation factor 2B from *Schizosaccharomyces pombe* (PDB:5B04) (Supplementary Fig. 4). For PDB:5ODC we could build a complete model of the protein with an r.m.s. deviation of 6.0 Å (TM-score 0.90)[30] starting from dimers. However, for PDB:5B04 it was not possible as the chains started overlapping when we tried to build a larger model. However, if we build trimers and then use all three dimers from these trimers we can build a complete model with an r.m.s. deviation of 7.3 Å (TM-score 0.86), showing that it is sometimes necessary to use larger subunits to assemble the complexes. Results from a follow-up study[31] show that it is often possible to build the structures of complexes if the subunits are well predicted. In summary, we find that it is possible to iteratively align structures of pairs of interacting proteins to build larger assemblies, but we also identified issues that limit this procedure at the moment.

Concluding discussion

We have predicted complex structures for pairs of human proteins known to physically interact from two different datasets based on different experimental approaches. We note that the source of data used for the protein interactions is important and impacts the fraction of models that can be confidently predicted. Our analysis suggests that protein interactions supported by a combination of affinity-, co-fraction- and complementation-based methods result in higher-confidence models. We believe these protein interactions tend to correspond to high-affinity interactions which are very likely to share a direct physical permanent interaction. We show that it is possible to use metrics from the models (for example, pDockQ score) to rank higher-confidence models, providing an additional accuracy level to large-scale PPI studies, and in the future to provide additional high-quality targets for detailed studies of stable complexes. Experimental data from crosslink mass spectrometry experiments provide an ideal resource for further validating these predictions via orthogonal means.

Based on comparisons with solved structures, we suggest that models with pDockQ > 0.5 are 80% likely to be correct. Additionally, models with lower scores (pDockQ > 0.23) are still 70% likely to contain many correct solutions and may highlight correct interfaces. Such lower-confidence models are likely to be useful for generating hypotheses and large-scale analyses of global properties. Equally important is the caveat that high-confidence predictions will still contain errors, and, in particular, we note that in protein complexes containing paralogous proteins (which is common in higher eukaryotes[32]), the current

procedure cannot identify the exact pairing of the protein. For such cases, additional methods need to be developed.

Structural models for protein interfaces are critical for understanding molecular mechanisms and the impact of mutations and post-translational modifications. We illustrate this using disease mutations and phosphorylation data. While much disease-associated variation is often found in noncoding regions of the genome, the growth of exome sequencing of large cohorts of patients will lead to discovering many more protein mutations linked to disease, which will require such large structural characteristics. Both for mutations and for phosphorylation sites, we think these analyses should be seen as generating hypotheses for further testing, and we make this information available in the supplementary material to facilitate such future work.

Finally, we show that it is in principle possible to build structural models for larger assemblies from predicted binary complexes. In a follow-up paper we have shown that it is possible to build large assemblies fully automatically by using predictions of dimers and trimers[31]. Aspects that may limit this include the structural homology between subunits, unknown subunit stoichiometries and limits in the predicted interactions[31]. Additional work will be needed to determine the exact stoichiometry and to design methods and score systems to build such larger complex assemblies, as well as to predict the interactions of proteins with weak and transient interactions.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-022-00910-8.

References

1. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
2. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
3. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.* **17**, e10016 (2021).
4. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2012).
5. Burley, S. K. et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).
6. Wang, X. et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
7. Kamburov, A. et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl Acad. Sci. USA* **112**, E5486–E5495 (2015).
8. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLoS Comput. Biol.* **11**, e1004518 (2015).
9. Beltrao, P. et al. Systematic functional prioritization of protein posttranslational modifications. *Cell* **150**, 413–425 (2012).
10. Nishi, H., Hashimoto, K. & Panchenko, A. R. Phosphorylation in protein-protein binding: effect on stability and function. *Structure* **19**, 1807–1815 (2011).
11. Šoštarić, N. et al. Effects of acetylation and phosphorylation on subunit interactions in three large eukaryotic complexes. *Mol. Cell. Proteom.* **17**, 2387–2401 (2018).
12. Betts, M. J. et al. Systematic identification of phosphorylation-mediated protein interaction switches. *PLoS Comput. Biol.* **13**, e1005462 (2017).
13. Zhang, Q. C. et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
14. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **42**, D374–D379 (2014).
15. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
16. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
17. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
19. Pozzati, G. et al. Limits and potential of combined folding and docking using PconsDock. *Bioinformatics* **38**, 954–961 (2022).
20. Akdel, M. et al. A structural biology community assessment of AlphaFold 2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
21. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
22. Humphreys, I. R. et al. Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
23. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
24. IMEx Consortium Curators et al. Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat. Commun.* **10**, 10 (2019).
25. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).
26. Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
27. Lawrence, R. T., Searle, B. C., Llovet, A. & Villén, J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat. Methods* **13**, 431–434 (2016).
28. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
29. Ochoa, D. et al. An atlas of human kinase regulation. *Mol. Syst. Biol.* **12**, 888 (2016).
30. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
31. Bryant, P. et al. Predicting the structure of large protein complexes using AlphaFold and sequential assembly. *Nat. Commun.* **13**, 6027 (2022).
32. Marchant, A. et al. The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *eLife* **8**, e46754 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format,

as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Protein interaction data and annotations

Human protein pairs known to physically interact were obtained from the hu.MAP dataset, retaining pairwise protein interactions with ≥0.5 confidence, and most interactions from the HuRI dataset. These interactions were further enriched by obtaining annotations on crosslinked peptides matched across pairs of interaction proteins, disease-related mutations and protein phosphorylation sites in the selected proteins. In addition, all nonhomologous pairs from 12 protein complexes (Supplementary Table 5) and 4,320 protein pairs from 2,102 different protein complexes (Supplementary Table 6) in CORUM[23] were used for additional analyses. A complete list of all datasets is available from the supplementary data. A subset of crosslink data was collected from refs. [33–43], and filtered for peptides unique to only one protein sequence. A crosslink was considered validated by the structure if the distance between the epsilon amino groups on the side-chains of the relevant pair of lysine residues was within 32 Å. Clinical missense variants associated with disease were collected from ClinVar. We selected only those having pathogenic or likely pathogenic effects, which were mapped to Uniprot protein sequences using VarMap. The final list of mutated positions was then compared with the interface positions. We obtained a list of protein phosphorylation sites with predicted functional relevance[26], phosphosite annotations[28] and regulation of phosphorylation sites across a large panel of conditions[29]. These phosphosites were also mapped to interface positions as defined by the predicted models. All protein interaction networks were processed using R packages igraph (v.1.2.5) and qgraph (v.1.9), and further graphical editing was done using Cytoscape[44].

Protein complex prediction

To predict protein complexes of pairwise interactions, we used the FoldDock pipeline[18] based on AlphaFold2 (ref. [17]). We used the option of fused + paired MSAs and ran the model configuration m1-10-1 as this provides the highest success rate accompanied by a 20-fold speed-up. Both the fused and paired MSAs were constructed by running HHblits on every single chain against Uniclust30. The fused MSA was generated by simply concatenating the output of each of the single-chain HHblits runs for two interacting chains. The paired MSA was constructed by combining the top hit for each matching OX identifier between two interacting chains, using the output from the single-chain HHblits runs.

pDockQ confidence score

To score models, we used features from the predicted complexes to calculate the predicted DockQ score, pDockQ. This score is defined with the following sigmoidal equation:

$$\mathrm{pDockQ} = \frac{0.707}{1 + e^{-0.03148(x - 388.06)}} + 0.03138$$

where

x = average interface plDDT*log(number of interface contacts).

The parameters were optimized to predict the DockQ score using the dataset from ref. [45]. The number of interface contacts is defined as elsewhere in this paper (any residues with an interface atom within 10 Å of the other chain), and the plDDT is the predicted lDDT score from AlphaFold2 taken over the interface residues as defined by the interface contacts.

Building larger complexes from binary protein interactions

A simple procedure to build larger complexes from a set of paired models was developed. All dimers in the set are by default ranked by their pDockQ values.

1. The building is started from a single dimer, by default the dimer with the highest pDockQ value. This is referred to as the 'complex'.
2. All other dimers in the set are then tried to be added to the 'complex'. Starting with the one with the second highest pDockQ, a chain is added to the complex if:
   (a) Exactly one chain of the dimer is identical to one chain in the complex
   (b) The structure of these two chains is similar enough (default TM-score > 0.8)
   (c) The dimer is then rotated so that the two chains overlap
   (d) The second chain in the dimer does not clash with more than 25% of its residues (Cα-Cα distance < 5 Å) with any chain in the complex.
3. If a chain is added, the procedure is started over again and repeated until no more chains can be added.

Analysis of phosphosites in the protein-protein interfaces

Phosphosite residues in interfaces were identified from a previously published comprehensive list of known human phosphosites[26]. Kinases associated with phosphorylation of interface residues were obtained from the PhosphositePlus database, and over-representation analysis of kinases was performed using a hyper-geometric test. Highly regulated interface phosphosites were defined as those with more than twofold change in phosphorylation in more than two perturbation conditions across a collated phosphoproteomics dataset comprising a range of physiological conditions and drug treatments[29]. Pearson correlation was calculated amongst these regulated phosphosites and clusters of co-regulated phosphosites were identified using hierarchical clustering ('Ward' method) of Euclidean distances of the correlation matrix. Phosphosite clusters were created by cutting the dendrogram at the appropriate level using the cutree ($h = 17$) function in R. Phosphosite clusters that were significantly regulated in each perturbation condition were identified by a Z-test from the comparison of fold changes in phosphosite measurements of all phosphosites in a cluster against the overall distribution of phosphorylation fold changes across the condition. Gene ontology over-representation of each cluster was performed separately using a hyper-geometric test in R. The gene ontology terms were obtained from the c5 category of the Molecular Signature Database (MSigDBv7.1)[46]. All over-representation analyses were performed using the enricher function of the clusterProfiler package (v.3.12.0)[6] in R.

Comparison with other databases

All proteins used here were mapped to UniProt[47] to retrieve subcellular localization, STRING[48] for coexpression and other interaction data, and gtex[49] for tissue-specific expression.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets and meta-data are available from https://doi.org/10.17044/scilifelab.16866202.v1. Further, all models generated as well as some of the multiple sequence alignments can be found at https://archive.bioinfo.se/huintaf2/. Source data are provided with this paper.

Code availability

All code used in this project can be found at https://gitlab.com/ElofssonLab/huintaf2/. Tools to run AlphaFold2 for combined folding and docking can be found at https://gitlab.com/ElofssonLab/FoldDock/.

References

33. Yugandhar, K. et al. MaXLinker: proteome-wide cross-link identifications with high specificity and sensitivity. *Mol. Cell. Proteom.* **19**, 554–568 (2020).

34. Schweppe, D. K. et al. XLinkDB 2.0: integrated, large-scale structural analysis of protein crosslinking data. *Bioinformatics* **32**, 2716–2718 (2016).
35. Klykov, O., van der Zwaan, C., Heck, A. J. R., Meijer, A. B. & Scheltema, R. A. Missing regions within the molecular architecture of human fibrin clots structurally resolved by XL-MS and integrative structural modeling. *Proc. Natl Acad. Sci. USA* **117**, 1976–1987 (2020).
36. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: an IMAC-enrichable cross-linking reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).
37. Klykov, O. et al. Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.* **13**, 2964–2990 (2018).
38. Fasci, D., van Ingen, H., Scheltema, R. A. & Heck, A. J. R. Histone interaction landscapes visualized by crosslinking mass spectrometry in intact cell nuclei. *Mol. Cell. Proteom.* **17**, 2018–2033 (2018).
39. Eliseev, B. et al. Structure of a human cap-dependent 48S translation pre-initiation complex. *Nucleic Acids Res.* **46**, 2678–2689 (2018).
40. Gestaut, D. et al. The chaperonin TRiC/CCT associates with prefoldin through a conserved electrostatic interface essential for cellular proteostasis. *Cell* **177**, 751–765.e15 (2019).
41. Klatt, F. et al. A precisely positioned MED12 activation helix stimulates CDK8 kinase activity. *Proc. Natl Acad. Sci. USA* **117**, 2894–2905 (2020).
42. Sabath, K. et al. INTS10-INTS13-INTS14 form a functional module of Integrator that binds nucleic acids and the cleavage module. *Nat. Commun.* **11**, 3422 (2020).
43. Mohamed, W. I. et al. The human GID complex engages two independent modules for substrate recruitment. *EMBO Rep.* **22**, e52981 (2021).
44. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
45. Green, A. G. et al. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1396 (2021).
46. Subramaniam, V., Vincent, I. R. & Jothy, S. Upregulation and dephosphorylation of cofilin: modulation by CD44 variant isoform in human colon cancer cells. *Exp. Mol. Pathol.* **79**, 187–193 (2005).
47. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
48. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
49. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

Acknowledgements

R.A.S. acknowledges funding through the European Union Horizon 2020 program INFRAIA project Epic-XS (project no. 823839) and the research program NWO TA with project no. 741.018.201, which is partly financed by the Dutch Research Council (NWO). A.E. was funded by the Vetenskapsrådet (grant no. 2016-03798 and 2021-03979) and the Knut and Alice Wallenberg Foundation. The computations/data handling were enabled by the supercomputing resource Berzelius provided by the National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg Foundation and SNIC, grant nos. SNIC 2021/5-297 and Berzelius-2021-29. J.E.B. acknowledges funding from the National Heart Lung and Blood Institute (grant no. 5R35GM13625) and the National Institute for General Medical Sciences (grant no. 5R01HL144778). P. Beltrao is supported by the Helmut Horten Stiftung and the ETH Zurich Foundation.

Author contributions

D.F.B. analyzed disease-causing mutations in interfaces, phosphosites and crosslinking with help from I.B.-H., D.M., A.S.D. and P. Beltrao. P. Bryant ran the prediction for the HuRI dataset. P.A., A.K., R.A.S., J.E.B. and A.L. provided data for the analysis. A.E. provided the prediction for the hu.MAP dataset and analyzed the complex structural features with help from P.K., G.P., A.S., P. Bryant and W.Z. P. Beltrao and A.E. wrote the manuscript with help from all authors.

Funding

Open access funding provided by Stockholm University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-022-00910-8.

Correspondence and requests for materials should be addressed to Petras Kundrotas, Pedro Beltrao or Arne Elofsson.

Peer review information *Nature Structural & Molecular Biology* thanks Shoshana Wodak and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Sara Osman, in collaboration with the *Nature Structural & Molecular Biology* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

A.12 Machine learning

Recommended reading for Chapter 12 (listed in Section 12.10).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Sapoval et al. (2022) — current progress and open challenges for deep learning across the biosciences (the basis of earlier exams).
  full text below · DOI · open access (CC BY 4.0)

  Current progress and open challenges for applying deep learning across the biosciences. Sapoval et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-29268-7

  > Deep Learning (DL) has recently enabled unprecedented advances in one of the grand challenges in computational biology: the half-century-old problem of protein structure prediction. In this paper we discuss recent advances, limitations, and future perspectives of DL on five broad areas: protein structure prediction, protein function prediction, genome engineering, systems biology and data integration, and phylogenetic inference. We discuss each application area and cover the main bottlenecks of DL approaches, such as training data, problem scope, and the ability to leverage existing DL architectures in new contexts. To conclude, we provide a summary of the subject-specific and general challenges for DL across the biosciences.

- Greener et al. (2021) — a guide to machine learning for biologists; how to choose and evaluate methods.
  DOI · via SU library (SU login)

  A guide to machine learning for biologists. Greener et al. (2021). *Nature Reviews Molecular Cell Biology*. doi:10.1038/s41580-021-00407-0

  > The expanding scale and inherent complexity of biological data have encouraged a growing use of machine learning in biology to build informative and predictive models of the underlying biological processes. All machine learning techniques fit models to data; however, the specific methods are quite varied and can at first glance seem bewildering. In this Review, we aim to provide readers with a gentle introduction to a few key machine learning techniques, including the most recently developed and widely used techniques involving deep neural networks. We describe how different techniques may be suited to specific types of biological data, and also discuss some best practices and points to consider when one is embarking on experiments involving machine learning. Some emerging directions in machine learning methodology are also discussed.

- LeCun et al. (2015) — deep learning, by three of its founders.
  DOI · via SU library (SU login)

  Deep learning. LeCun et al. (2015). *Nature*. doi:10.1038/nature14539

  > Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

- Krogh (2008) — what artificial neural networks are, in three pages.
  DOI · via SU library (SU login)

  What are artificial neural networks?. Krogh (2008). *Nature Biotechnology*. doi:10.1038/nbt1386

> Artificial neural networks have been applied to problems ranging from speech recognition to prediction of protein secondary structure, classification of cancers and gene prediction. How do they work and what might they be good for?

- Walsh et al. (2021) — DOME; what a supervised machine learning study in biology should report.
  DOI · via SU library (SU login)

  DOME: recommendations for supervised machine learning validation in biology. Walsh et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01205-4

- Chicco and Jurman (2020) — why the Matthews correlation coefficient is better than F1 and accuracy.
  full text below · DOI · open access (CC BY 4.0)

  The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. Chicco & Jurman (2020). *BMC Genomics*. doi:10.1186/s12864-019-6413-7

  > Background To evaluate binary classifications and their confusion matrices, scientific researchers can employ several statistical rates, accordingly to the goal of the experiment they are investigating. Despite being a crucial issue in machine learning, no widespread consensus has been reached on a unified elective chosen measure yet. Accuracy and F 1 score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification tasks. However, these statistical measures can dangerously show overoptimistic inflated results, especially on imbalanced datasets. Results The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. Conclusions In this article, we show how MCC produces a more informative and truthful score in evaluating binary classifications than accuracy and F 1 score, by first explaining the mathematical properties, and then the asset of MCC in six synthetic use cases and in a real genomics scenario. We believe that the Matthews correlation coefficient should be preferred to accuracy and F 1 score in evaluating binary classification tasks by all scientific communities.

The open-access papers are reproduced below under their original licences.

nature COMMUNICATIONS

REVIEW ARTICLE

https://doi.org/10.1038/s41467-022-29268-7 OPEN

Current progress and open challenges for applying deep learning across the biosciences

Nicolae Sapoval[1,6], Amirali Aghazadeh[2,6], Michael G. Nute[1], Dinler A. Antunes[3], Advait Balaji[1], Richard Baraniuk[4], C. J. Barberan[4], Ruth Dannenfelser[1], Chen Dun[1], Mohammadamin Edrisi[1], R. A. Leo Elworth[1], Bryce Kille[1], Anastasios Kyrillidis[1], Luay Nakhleh[1], Cameron R. Wolfe[1], Zhi Yan[1], Vicky Yao[1] & Todd J. Treangen[1,5]✉

Deep Learning (DL) has recently enabled unprecedented advances in one of the grand challenges in computational biology: the half-century-old problem of protein structure prediction. In this paper we discuss recent advances, limitations, and future perspectives of DL on five broad areas: protein structure prediction, protein function prediction, genome engineering, systems biology and data integration, and phylogenetic inference. We discuss each application area and cover the main bottlenecks of DL approaches, such as training data, problem scope, and the ability to leverage existing DL architectures in new contexts. To conclude, we provide a summary of the subject-specific and general challenges for DL across the biosciences.

The recent success of AlphaFold2[1] in predicting the 3D structure of proteins from their sequences highlights one of the most effective applications of deep learning in computational biology to date. Deep learning (DL) allows for finding a representation of the data with multiple layers of abstraction using complex models that are composed of several layers of nonlinear computational units (Fig. 1). Observed through the success of DL in a broad variety of application domains, the efficacy of using DL depends on the development of specialized neural network architectures that can capture important properties of the data such as spatial locality (convolutional neural networks – CNNs), sequential nature (recurrent neural networks – RNNs), context dependence (Transformers), and data distribution (autoencoders – AEs). Figure 1 illustrates six DL architectures that have found the most applications within the realm of computational biology. We refer the reader to LeCun et al.[2] for a complete review of DL methods and architectures and keep the focus of the paper on computational biology applications. These DL models have revolutionized speech recognition, visual object recognition, and object detection and have lately played a key role in solving important problems in computational biology. The applications of DL in other areas of computational biology, such as functional biology, are only growing while other areas, such as phylogenetics, are in their infancy. Given the wide divide between the receptiveness of DL in different areas in computational biology,

[1] Department of Computer Science, Rice University, Houston, TX, USA. [2] Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, USA. [3] Department of Biology and Biochemistry, University of Houston, Houston, TX, USA. [4] Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [5] Department of Bioengineering, Rice University, Houston, TX, USA. [6] These authors contributed equally: Nicolae Sapoval, Amirali Aghazadeh. ✉email: treangen@rice.edu

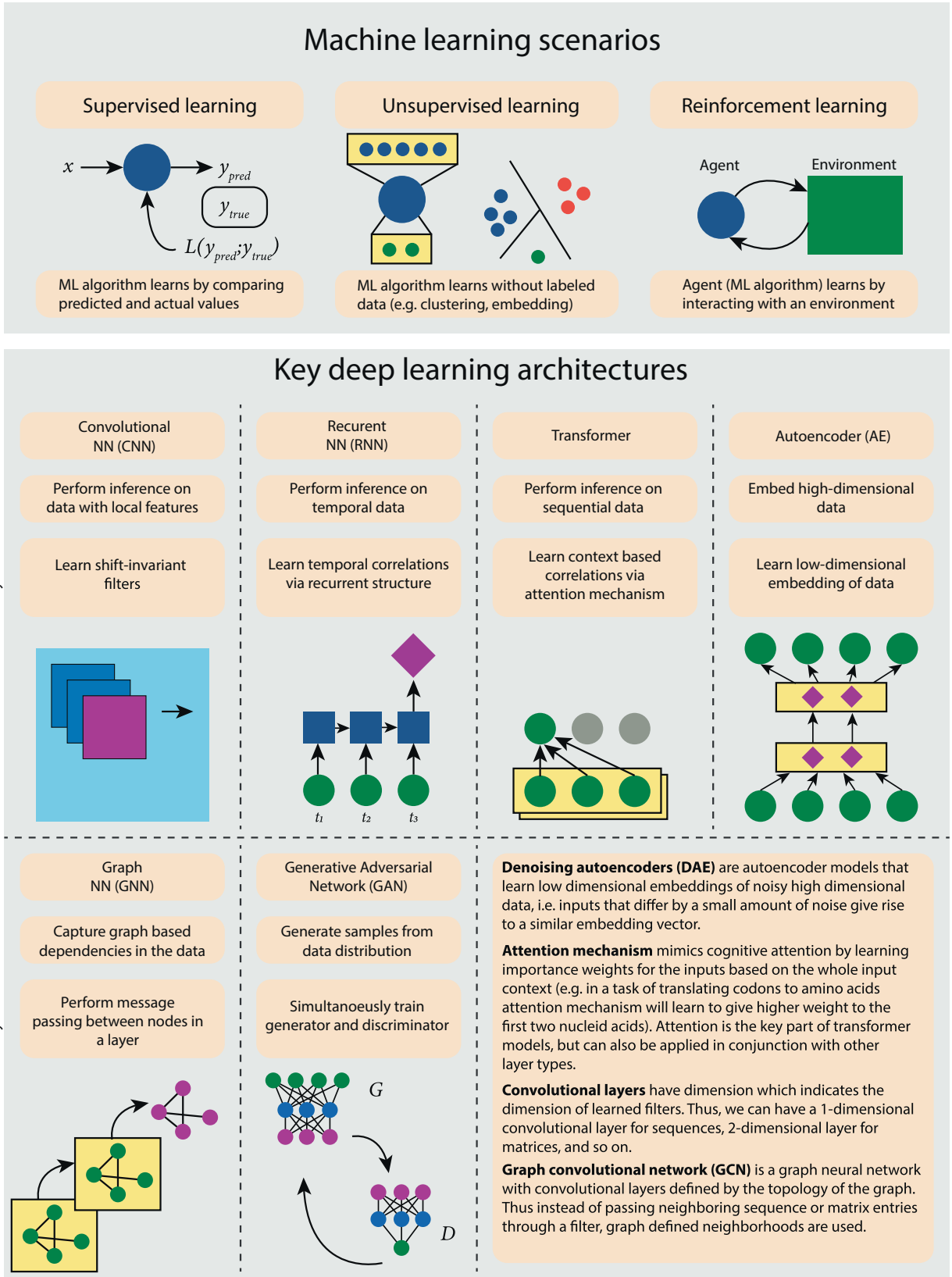


Fig. 1 An overview of machine learning scenarios and commonly used DL architectures. Top panel encapsulates the three most common paradigms of machine learning: supervised learning in which dataset contains ground truth labels, unsupervised learning in which dataset does not contain ground truth labels, and reinforcement learning in which an algorithmic agent interacts with a real or simulated environment. The bottom panels provide an overview of the most prevalent DL architecture ideas each designed to achieve specific highlighted goals. An additional set of short descriptions is provided for other common components of DL architectures mentioned in the manuscript.

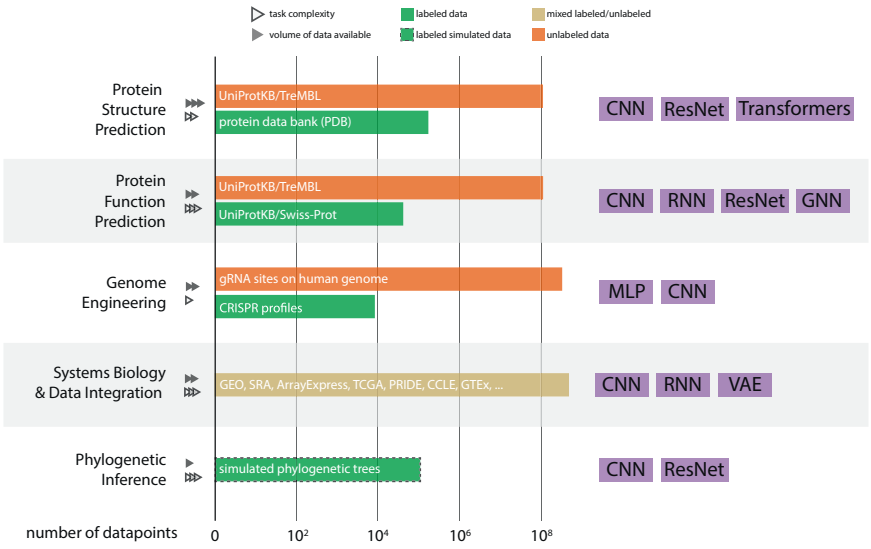


Fig. 2 A summary view of the major labeled and unlabeled datasets, and the architectures being used in deep-learning methods in computational biology. For each of the areas considered in this manuscript, it summarizes estimated sizes of key datasets and databases, as well as the projected growth rate of these. Additionally the rightmost column summarizes the most popular DL architectures applied to the corresponding areas in biosciences.

Table 1 Impact of Deep Learning on Computational Biology.

	Protein structure prediction	Protein function prediction	Genome engineering	Systems biology and data integration	Phylogenetic inference
Paradigm shifting	✓				
Major success		✓	✓		
Moderate success				✓	
Minor success					✓

Each of the subareas in biosciences considered in this manuscript is assigned a level of success of the DL applications based on the relative performance of DL as compared to other ML and classical methods.

some key questions remain unanswered: (1) What makes an area prime for DL methods? (2) What are the potential limitations of DL for computational biology applications? (3) Which DL model is most appropriate for a specific application area in computational biology?

In this paper, we aim to address these foundational questions from the lens of computational biology. The answers, however, are highly task specific and can only be addressed in the context of the corresponding applications. The pitfalls of applying machine learning (ML) in genomics have been discussed in Whalen et al.[3], but our goal is to provide a perspective on the impact of DL across five distinct areas. While there are multiple areas of interest in the biosciences where DL has achieved notable successes (e.g. DeepVariant[4], DeepArg[5], metagenomic binning[6], and lab-of-origin attribution[7]), we aim to only focus on a few diverse and broad subtopics. In those areas we evaluate the improvements that DL has had over classical ML techniques in computational biology with varying levels of success to date (Fig. 2). For each area, we explore limitations of current approaches and opportunities for improvement, and include practical tips. We anchor our discussions around five broad, distinct areas in computational biology: protein structure prediction, protein function prediction, genome engineering, systems biology and data integration, and phylogenetic inference (Table 1). These areas provide a range of impact levels from major paradigm shifts (AlphaFold2) to DL applications in their infancy (phylogenetic inference); and collectively, they provide rich enough technical diversity to address the questions raised in this perspective. Over the next several subsections, we will review progress in each of the four computational biology topics, ordered from (i) paradigm shifting (where DL clearly outperforms other ML and classical approaches, and provides a field-wide impact), (ii) major success (where DL performance is typically higher than of that of other ML and classical approaches), (iii) moderate success (where DL performance is typically comparable to other ML and classical approaches) to (iv) minor successes (where DL methods are not widely adopted or underperform compared to other ML and classical approaches), and then discuss common challenges for DL in biosciences (Table 2).

Table 2 Commonly faced challenges in computational biology and potential solution avenues when using DL.

Challenge	Experimental/non-DL solution	DL solution
Biased results	Improve study design	Identify forms and sources of technical bias Fair AI approaches
High infrastructure costs	Optimize code performance Parallelize code Sub-sample analyzed data	Optimize DL architecture Parallelize to low-cost devices Condense training data (e.g. coresets)
Lack of explainability	Statistical analyses	Explainable post-hoc methods
Limited training data	Generate and label more data	Data augmentation (e.g. GANs)
Overfitting	Regularization	Dropout Early stopping Smaller models Additional training data
Poor performance on novel data	Expand databases	Use larger models Analyze generalization potential

Paradigm shifting successes of DL

Protein structure prediction. We start our discussion with protein structure prediction which is arguably one of the most successful applications of DL in computational biology; this success is what we refer to as a paradigm shift. It is largely known that the protein's amino acid sequence determines its 3D structure, which is in turn directly related to its function (e.g., chemical reaction catalysis, signal transduction, scaffold, etc.)[8,9]. The history of protein structure prediction problem goes back to the determination of the 3D structure of myoglobin by John Kendrew in the 1950s which was a landmark in biochemistry and structural biology[10]. Since then, X-ray crystallography has become the gold-standard experimental method for protein structure determination[11,12], as well as the reference to validate computational models for protein structure prediction. Considering the high cost and technical limitations of X-ray crystallography, and the growing access to biological sequences following the Human Genome Project, predicting the 3D structure of a protein from its sequence became the Mount Everest in computational biology[8]; a challenge broadly known as the "protein folding problem". Initial efforts concentrated on the use of biophysically-accurate energy functions and knowledge-based statistical reasoning, but faster progress was recently achieved with a greater focus on DL.

One of the key reasons for the recent success of DL in this area has been the wealth of unsupervised data in the form of multiple sequence alignment (MSA)[1,9,13–17], which has enabled learning a nonlinear evolution-informed representation of proteins. Progress in the field has been accelerated by the creation of a bi-annual international competition, called the Critical Assessment of Protein Structure Prediction (CASP). Launched in 1994, CASP created the means to objectively test available methods through blind predictions, providing competing groups with a set of challenges (i.e., sequences of proteins with unknown structures), and evaluating their performances against the respective experimentally-determined structures. In their first participation in CASP13, AlphaFold, implemented by DeepMind group at Google, made the news by clearly outperforming the second best method[14], and nearly twice beyond the projection based on previous editions[18]. Following recent trends in the field[13,16,19,20], AlphaFold and AlphaFold2 leverage the combined use of DL and MSA[18,21]. This proved to be a winning strategy which was able to overcome the lack of large training datasets on protein structure. The Protein Data Bank (PDB)[22] is the reference database for experimentally-determined macromolecular structures, and currently hosts close to 180,000 entries. This is a small number of data points for a complex mapping involved in the problem, and these are further biased by technical constraints of the experimental methods. Protein sequence data, on the other hand, is available on a much larger scale. Therefore, MSA allows modeling methods to extract pairwise evolutionary correlations from this larger corpus of data, maximizing the learning on available structural data. Other key factors for the success of DL in this area include innovation in model design such as new attention strategies tuned towards invariances and symmetries in proteins, graph-based representations, and model recycling strategies.

The impact of AlphaFold2 on the field of structural biology is undeniable; it successfully demonstrated the use of a DL-based implementation for high accuracy protein structure prediction[21]. This achievement is already driving and accelerating further developments in the field, as highlighted by the remarkable number of early citations. In addition, DeepMind has partnered with the European Molecular Biology Laboratory (EMBL)[23] to create an open-access database of protein structures modeled with AlphaFold2[17]. The database already covers 98.5% of human proteins, for which at least 36% of the amino acid residues were predicted with high confidence. Finally, rather than retiring experimental methods, DL-based methods might augment the accuracy and reach of experimental methods as demonstrated by preliminary applications to solving challenging structures with data from X-ray crystallography and cryo-EM[1,15]. However, many caveats, limitations and open questions[8,9] remain. In particular, while AlphaFold2 successfully predicts the static structure of a protein, many key insights about protein's biological function come from its dynamic conformations. Furthermore, dynamics of interaction of multiple proteins still present open challenges in the field. Moving forward, it will be important to monitor the application of DL to these follow up research areas.

Major successes of DL

Protein function prediction. Predicting protein function is a natural next step after protein structure prediction. Protein function prediction involves mapping target proteins to curated ontologies, such as Gene Ontology (GO) terms, Biological Processes (BP), Molecular Functions (MF) and Cellular components (CC). Protein structure can convey a lot of information about these ontologies, however, there is no direct mapping between the two and the mapping is often very complex[24]. Despite the tremendous growth of protein sequences available in the UniProtKB database, functional annotations for the vast majority of proteins still remain partly or completely unknown[25]. Limited and imbalanced training examples, a large output space of possible functions and the hierarchical nature of the GO labels are some of the main bottlenecks associated with functional annotation of proteins[26]. To overcome some of the issues recent methods have leveraged features from different sources including sequence[27], structure[22],

interaction networks[28], scientific literature, homologies, domain information[29] and even incorporate one or a combination of DL architectures to handle different stages of prediction task (e.g. feature representation, feature selection, and classification).

One of the most successful DL approaches to the problem, DeepGO[30] incorporated CNN to learn sequence-level embeddings and combines it with knowledge graph embeddings for each protein obtained[31] from Protein-Protein Interaction (PPI) networks. DeepGO was one of the first DL based models to perform better than BLAST[32] and previous methods on functional annotation tasks over the three GO categories[30]. An improved version of the tool, DeepGOPlus[33] emerged as one of the top performers when compared to other tools in the CAFA3 challenge across the three GO categories[33]. DeepGOPlus used convolutional filters of different sizes with individual max-pooling to learn dense feature representations of protein sequences embedded in a one-hot encoding scheme. The authors showed that combining the outputs from CNN with homology-based predictions from DIAMOND[34] can result in better predictive accuracy.

Unsupervised methods such as DAEs also have been instrumental by learning dense, robust, and low-dimensional representations of proteins. Chicco et al.[35] developed a DAE to represent proteins for assigning missing GO annotations and showed 6% to 36% improvements compared to non-DL methods over six different GO datasets. Miranda and Hu[36] introduced the Stacked Denoising Autoencoders (sdAE) to learn more robust representation of proteins. Gilgorijevic et al. introduced deepNF[37] that uses Multimodal DAE (MDA) to extract features from multiple heterogeneous interaction networks which outperform methods based on matrix factorization and linear regression [37]. Methods for learning low-dimensional embeddings of proteins continue to grow.

Beyond just predicting Gene Ontology labels, studies have also focused on several other task-specific functional categories such as identifying specific enzyme functions[38] and potential post-translational modification sites[39]. These studies are a fundamental step towards developing novel proteins with specialized functions or modifying the efficacy of existing proteins as seen in the recent advances of DL in enzyme engineering[40]. Going forward, applications of deep learning in engineering proteins tailored to specific functions can help increase throughput of candidate proteins for pharmaceutical applications among other domains.

Besides these canonical architectures, there have been other approaches that have used a combination of the above methods for function classification[41]. Overall, previous results indicate that models integrating features from multi-modal data types (e.g., sequence, structure, PPI, etc) are more likely to outperform the ones that rely on a single datatype. Trends from literature indicate that relying on task-specific architectures could help greatly enhance the feature representation from respective data types. Future work in this direction could focus on combining DAEs and RNNs for sequence based representation, and Graph Convolutional Networks (GCNs) for structure based as well as PPI based information. Combining these representations in a hierarchical classifier such as the multi-task DNN with biologically-relevant regularization methods[42,43] could allow for an explainable and computationally feasible DL architecture for protein function prediction.

Genome engineering. Biomedical engineering, and in particular genome engineering, is an important area in biology where DL models have been increasingly employed. Among genome engineering technologies, clustered regularly interspaced short palindromic repeats (CRISPR), i.e., a family of DNA sequences found in the genomes of prokaryotic organisms, have been recently used as a guide to recognize and cleave specific locations on the human genome. In the CRISPR-associated protein 9 (Cas9) technology, a single-guide RNA (gRNA) steers the protein to a specific genomic target. When the 20-nucleotide gRNA sequence complements the genome, Cas9 creates a double-strand break (DSB) on the targets (an on-target event). Due to the ability to precisely target specific locations on the genome, we have observed enormous advancements in CRISPR-based editing technologies since the development of Cas9. However, recent studies have shown that multiple mismatches between the gRNA and the genomic targets are tolerated and, as a result, Cas9 can cut unwanted locations on the genome (an off-target event). Off-target edits have pathogenic effects on the functionality and integrity of the cell. Therefore, the full clinical deployment of Cas9 has been slow due to the insufficient efficiency, reliability, and controllability challenges for therapeutic purposes. As a result, reducing off-target while improving the on-target efficiency has been an important ultimate goal in genome engineering target by DL techniques.

The sheer complexity of the biological process involved in modeling the DNA repair process and the growing availability of labeled data caused by a rapid drop in the cost of CRISPR assays, have made DL-based methods particularly successful choices to find the root cause of these inefficiencies. The use of DL models was triggered by the observation that the on-target and off-target events and the DNA repair outcome[44] are predictable by the sequence around the DSB, its location on the genome, and the potential mistargeted sequences on the genome. Several computational tools have been successfully developed to design gRNAs with maximum on-target activity and minimum off-target effects[45]. DeepCas9 is among CNN-based models which learns functional gRNAs directly from their canonical sequence representation[46,47]. The success of DeepCRISPR, on the other hand, relies on extracting about half a billion unlabeled gRNA sequences from the human coding and non-coding regions and learning a low-dimensional representation of the gRNA[48]. DeepCRISPR also uses a data augmentation method to create less than a million sgRNAs with known knockout efficiencies to train a larger CNN model. CnnCrispr uses a language processing model to learn the representation of gRNA and then employs a combination of bidirectional LSTM and CNN[49] while RNNs have been the reason for the success of other models[50]. Attention mechanism has also been shown to improve the accuracy in predicting on and off target effects[51,52]. ADAPT[53] is another recent CNN-based method for fully-automated CRISPR design for vertebrate-infecting viral diagnostics which owes its success to the construction of a massive training CRISPR dataset. Recent methods for predicting the DNA repair outcome employ other strategies. SPROUT compensates the lack of labels on harder-to-collect human CD4+ T cells by predicting a summary statistics of the DNA repair outcome[54]. FORECasT employs a larger dataset from easier-to-collect human chronic myelogenous leukemia cell-line (K562)[55]. InDelphi creates hand-designed features of the input sequence including the length and GC content of the homologous sequences around the cut site[56] while CROTON avoids feature engineering and instead performs neural architecture search[57]. All these strategies help reducing the number of labeled data points required to learn the input-output mapping.

The future of DL is geared towards new editing technologies such as CRISPR-Cas12a (cpf1)[58], base editors[59], and prime editors[60]. While these methods do not introduce DSBs, their efficiency is still improving[61]; in fact, DL has already shown promise in predicting the efficiency of Adenine base editors (ABEs) and Cytosine base editors (CBEs)[59] as well as prime editor 2 (PE2)

activities in human cells[60]. The future challenges, however, are in understanding these models. CRISPRLand is a recent framework which takes the first step towards interpretation and visualization of DL models in terms of higher-order interactions[62]. Besides explainablity, we speculate that methods that enable an uncertainty estimate of the prediction outcome become more prevalent in genome editing. Further, due to the significant cell-type effects on the efficiency of the CRISPR experiments, it is critical to be aware of the distribution shifts in deploying DL models in genome engineering. The integration of domain adaptation[63] methods to limit the effect of such distribution shifts are among other important future directions.

Moderate successes of DL

Systems biology and data integration. Systems biology takes a holistic view of modeling complex biological processes to ultimately unravel the link between genotype and phenotype. Integration of diverse -omics data is central in bridging this gap, enabling robust predictive models that have led to several recent breakthroughs, spanning from basic biology[64] to precision medicine. These data are now more accessible than ever, due to improvements in sequencing technologies and the establishment of open access public repositories where researchers can deposit their own studies, such as SRA[65], GEO[65], ArrayExpress[66], and PRIDE[67]; and large coordinated efforts with structured multi-omic datasets: TCGA[68], CCLE[69], GTEx[70], and ENCODE[71]. Given recent successes and the prevalence of both single and co-assay data, the field is now focused on integrating different data types (e.g., genomics, transcriptomics, epigenomics, proteomics, metabolomics) on single individuals, across many individuals, within and between phenotypic groups, and across different organisms. Data integration tasks fall into two main categories: 1) integration across different platforms and studies of a single data type, at times with other non-omics data (e.g., protein-protein interactions, pathway annotations, motif presence) and 2) integration between different -omic data types (e.g., RNA-seq, ChIP-seq, ATAC-seq, BS-seq). Much progress has been made on integration within a single data type, especially transcriptomics data, with classical ML and statistical approaches developed for batch correction[72–75], modeling global gene co-expression patterns[76], Bayesian integration strategies for function prediction[77,78], and phenotype classification[79]. More recently, the increasing prevalence of single-cell transcriptomics has given rise to a new host of classic ML[80–82] and DL[83,84] approaches for data integration across experiments. DL methods in this space have arisen out of the need for methods that scale well with the large number of cells and ability to model non-linear patterns of cell similarity[83,85]. Here, we have only skimmed the surface of methods being developed for expression data, but this trend is emerging for other -omics data types, similarly driven by the resolution of improved high-resolution experimental assays[86,87]. Broadly, data integration analyses that simultaneously combine data types together, either from different studies or different types, typically fall into one of three categories, given the stage at which the integration is performed[88]: concatenation-based, transformation-based, or model-based. While data integration across studies can be data of the same type, here we focus on methods that specifically integrate across different -omics types, as these questions introduce additional technical challenges and complexity.

Concatenation-based integration methods perform data integration early in the method pipeline by combining data, in raw or processed forms, before any joint modeling and analysis. Traditional ML concatenation-based methods are often unsupervised and typically use automatic feature extraction techniques such as lasso[89], joint clustering schemes[90], and dimensionality reduction[91] to find relevant signal. These methods are usually applied to well-curated, multi-omic datasets from large consortia (e.g., TCGA), and thus most often have been used to find meaningful patient subgroups characterized by distinct patterns across data modalities. More recently, autoencoders have been used as an initial data processing step to generate lower dimensional embeddings that are then concatenated together as features for downstream models[92,93]. These approaches have improved performance over existing methods likely due to the advantages autoencoders have in denoising tasks, as well as their abilities to model nonlinear latent structure, even without sample labels.

Instead of directly concatenating separate latent embeddings, some groups have pursued transformation-based integration methods by modeling data jointly by mapping to a common representation (e.g., graph or kernel matrix). Historically, classic transformation-based ML methods use known anchor references[94], kernel[95], or manifold methods[96] to align multi-omics data. This is a rapidly growing area in data integration, especially for DL methods. Building off of the use of anchors from classical ML methods, new state-of-the-art methods frequently train single modality autoencoders, followed by an alignment procedure across modalities[97]. This direction is exciting, because once trained, the models can be used to predict an unobserved modality given a single data type. Additional exciting developments harness the power of these embedding representations together with other DL methods, including CNNs and RNNs for wide ranging predictive tasks, including cell fate[98], drug response[99], survival[92,100], and clinical disease features[101].

Perhaps the most straightforward way to integrate multi-modal data is to train individual data modality models, then integrate them by combining the results from the individual models, termed model-based integration. To some degree, this is similar to ensemble approaches frequently used in classical ML. Methods in this space can take wide-ranging approaches, including building data modality-specific networks before fusing them using message-passing theory[102] or combining different data representations using a discriminative learning approach[103]. DL methods have yet to gain much momentum for model-based integration, likely because the very nature of most DL methods blurs the line between the transformation-based and model-based paradigms. Classical approaches here try to bridge data modalities by finding a common modeling space, while DL naturally can identify common representations and model them jointly, thus circumventing the need for separate modeling and integration steps. While it is clear that deep neural networks will likely lead to better performance in data integration tasks, it is also important to keep in mind the limitations of DL, as well as important areas for continued research. Specifically, it is known that DL has the tendency to overfit to data. On the other hand, in data integration tasks, batch effects can be prevalent and it is often easy to have "contamination" between the training and test sets, all of which can lead to inflated performance estimates. Thus, it is important to carefully set up truly independent evaluation sets and identify appropriate performance baselines[3]. Furthermore, while genome-wide and whole transcriptomics datasets have broad coverage across the genome and transcriptome, human data (and in some cases, model organism data) is often skewed towards a disproportional amount of sick individuals[104], is sex-biased towards men[105], and biased by race with an over-represented population of Europeans[106]. These biases can result in spurious associations that plague all ML methods, but may be particularly difficult to identify when using DL.

Minor successes of DL

Phylogenetics. A phylogeny is an evolutionary tree that models the evolutionary history of a set of taxa. The phylogeny inference problem concerns building a phylogeny from data—often

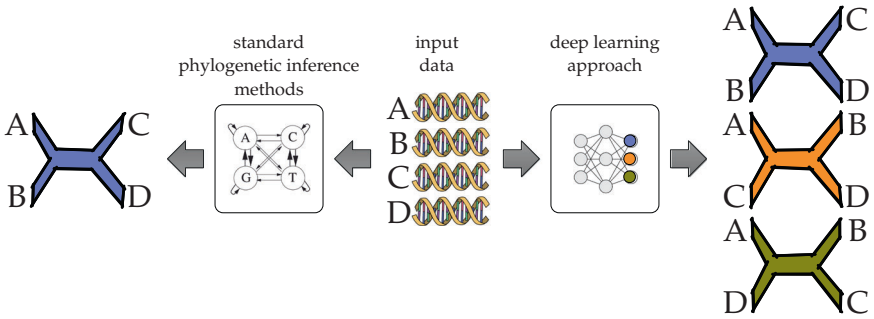


Fig. 3 Standard and DL approaches to phylogenetic inference. The input consists of sequences (DNA sequences in this illustration) obtained from the taxa of interest. Here, the taxa are **A, B**, **C**, and **D**. In standard approaches, such as maximum likelihood and maximum parsimony, a generative model in the form of a tree whose leaves are labeled by the four taxa is inferred. In the recently introduced DL approach to phylogenetic inference, the problem is viewed as a classification task where the network outputs correspond to the three possible tree topologies whose leaves are labeled by the taxa **A**, **B**, **C**, and **D**.

molecular sequences—obtained from the set of taxa under investigation [107]. Figure 3 illustrates the phylogeny inference problem on four taxa; in this case it can be viewed as a classification problem among three possible topologies. However, classification methods have a major limitation in that they cannot infer branch lengths, nor do they scale beyond a very small number of taxa because the number of possible topologies (classes) grows super-exponentially with this variable. But perhaps more importantly, classifiers like DL models require training data, and benchmark data where the true phylogeny is known is almost impossible to obtain in this field. Instead, simulations have been the method of choice for generating training data, but this is a major dependency and methods are known to have divergent performance on simulated and biological data[108]. For complex versions of the phylogeny inference problem, more realistic simulation protocols are needed. Finally, phylogenetic inference on a single gene is in one sense a simplified problem itself: inferring a single phylogeny from genome-wide data introduces the complication that different genes can have different histories, or the true phylogeny might be a network[109], rather than a tree. For these reasons DL has either had limited success or been restricted to small sub-problems aside from the main inference task.

Nonetheless, there have been attempts to use DL for the classification task as described above. The Self-Organizing Tree (SOTA) algorithm[110] is a two-decades old unsupervised hierarchical clustering method based on a neural network to classify sequences and reconstruct phylogenetic trees from sequence data. SOTA follows the SOM (Self-Organizing Map) algorithm in growing cell structures from top to bottom dynamically until a desired (user-provided) taxonomic level is reached. Recently CNNs have been used to infer the unrooted phylogenetic tree on four taxa (called a quartet)[111,112]. Authors used simulated data for training a classifier which assigns sequences to their phylogenetic tree (Fig. 3). But an analysis of the performance of the method of Zou et al.[112] by Zaharias et al.[113] shows that CNNs were not as accurate as other standard tree estimation methods, e.g., maximum likelihood, maximum parsimony, and neighbor joining, neither in terms of quartet estimation nor in terms of full tree estimation, especially when the sequence length was relatively short and/or rates of evolution were not sufficiently low. A potential workaround is to approach phylogeny inference as a graph generation problem, a more complex learning task.

Distance-based methods are another class of commonly used techniques for phylogenetic inference among which the neighbor joining method is the most common one, and DL has been applied to improve the distance representation. Jiang et al.[114] addressed the phylogenetic placement problem, i.e., the problem of adding a new taxon to a given tree without having to rebuild the tree from scratch, by training a CNN using a simulated backbone tree and sequences. Given the backbone tree with its associated and query sequences, the model outputs an embedding of the query and reference species which can be used as input to some distance-based phylogenetic placement tools, which then places the query sequences onto the reference tree. Bhattacharjee et al.[115] addressed the data imputation problem in the incomplete distance matrix using autoencoders. However, the key limitation of these methods is that trees cannot be reliably embedded into a Euclidean space of low dimensions[116]. Hyperbolic space, on the other hand, has been demonstrated to be more suitable for representing data with hierarchical latent structure[117].

Other applications have used DL to aid in a more traditional inference pipeline. For example, the particular likelihood model to use for a maximum-likelihood search is often taken for granted as user decision, but a recent method used DL to optimize this decision[118]. In another case, DL was used to aid decision-making in the tree-search algorithm used in a traditional maximum likelihood heuristic. Finally, a very recent application uses a sparse learning model for something almost like the reverse process: given a phylogeny, it identifies the portions of a genome that most directly explain or relate to that model[119]. This can be used to validate phylogenetic inference as well as guide downstream analyses such as hypothesis generation and testing.

A traditional problem is the inference of perfect phylogeny where every site in the sequences mutates at most once along the branches of the tree. The problem of determining whether a perfect phylogeny exists and inferring it, if one exists, from binary data that is assumed to be correct is polynomially solvable. However, if the data is assumed to have errors, one approach to inferring a perfect phylogeny is by solving the minimum-flip problem: given a binary matrix of mutations - where each entry represents the presence (state 1) or absence (state 0) of mutation in a sample and a site - that does not admit a perfect phylogeny, the minimum number of "state flips" (from 0 to 1 or 1 to 0) to the data is sought so that a perfect phylogeny is admitted. Sadeqi Azer et al.[120] used an existing DL framework originally designed for solving the traveling salesman problem to tackle this problem[121]. Here, the input consists of the inferred single-nucleotide variations (SNVs) in single cells across different sites. The output is a matrix that admits a perfect phylogeny with the minimum number of state flips from the input matrix. The input matrix is flattened and passed through convolutional layers for encoding. The encoded data is fed to a Long Short Term Memory (LSTM) layer as a decoder. Then, an attention layer takes the outputs of the LSTM layer to score the entries of the mutation matrix according to the impact that flipping them might have on minimizing the overall number of state flips. This architecture

results in a probability distribution on the entries of the input matrix that is used for flipping them. The model is trained using simulated data where the matrix and the number of flips to perform are provided. The key limitation of this approach is that there is no guarantee that the output admits a perfect phylogeny because the cost function might not be fully optimized.

Taken altogether, these related successes are impressive, but given the challenges outlined above it is difficult to conceive of an end-to-end DL model to directly estimate phylogenetic trees from raw data in the near future. And if one were to be developed, given its reliance on (likely simulated) training data, its applicability to actual biological sequences will need to be carefully validated before traditional phylogenetic methods are displaced.

General challenges for DL in the biosciences

Not all applications of DL have been equally successful in computational biology. While in some areas such as protein structure prediction and genome editing DL has found major success, in other areas like phylogenetic inference, DL has faced major hurdles (Table 1). Most common issues faced by DL approaches stem from the lack of annotated data, inherent absence of the ground truth for non-simulated datasets, severe discrepancies between training data distribution and real-world test (e.g., clinical) data distribution, potential difficulties in result benchmarking and interpretation, and finally overcoming the biases and ethical issues in datasets and models. Additionally, with the growth of the data and DL models, training efficiency has become a major bottleneck for progress.

Specifically, the success of DL in different subareas in computational biology highly relies on the availability and diversity of standardized supervised and unsupervised datasets, ML benchmarks with clear biological impact, the computational nature of the problem, and the software engineering infrastructure to train the DL models. The remaining challenges of DL in computational biology are tied with improving model explainability, extracting actionable and human-understandable insights, boosting the efficiency and limiting the training costs, and finally mitigating the growing ethical issues of DL models; innovative solutions are emerging in DL and computational biology communities (Table 2). We will now review two key areas for improvement: (i) Explainability and (ii) Training efficiency.

Explainability. Perhaps one of the most critical limitations of DL models today, especially for biological and clinical applications, is that they are not as explainable as the simpler regression models in statistics; it is challenging to explain what each node of the network represents and how important it is to model performance. The highly nonlinear decision boundaries of DNNs and their overparameterized nature, which enable them to achieve high prediction accuracy, make them hard to explain as well. This lack of explanability becomes an important issue in computational biology, because trustworthiness of DNNs is arguably one of the most pressing problems in biological and sensitive clinical decision making applications. In fact, in biology often the question of why a model can predict well is as important as how accurately it can predict a phenomenon. For example in protein structure/function prediction we would like to know what rules in a predictive model govern the 3D geometry of a protein and its properties; in genome editing we aim to understand the biological DNA repair processes inferred from CRISPR models; in systems biology we aim to know the specific molecular differences that give rise to different phenotypes; in phylogenetics we aim to know the features that enable us to infer a phylogenetic tree. Addressing these questions are key in producing biological knowledge and creating actionable decisions in the clinical settings.

There have been some efforts in the ML community to develop methods to explain "black-box" DL models in the past few years[122]. Earlier works were developed in computer vision and biomedical applications, some of which have been applied to problems in computational biology as well. Activation maximization is a large class of algorithms which searches for an input which maximizes the model response typically by using gradient descent[123,124]; the idea is to generate an input that best symbolizes an outcome. To make them human-interpretable, the input gets regularized using closed-form density functions of the data or GANs that mimic the data distribution. Methods that address the explainability question use more direct ways to gain insights from the NN function using their Taylor expansion[125] or Fourier transform[42,62]. The explanation takes the form of a heatmap which shows the importance of each input feature. Sensitivity analysis is another popular method of this sort which finds the input features to which the output is most sensitive to using backpropagation[126]; this has been used for classification and diagnostic prediction of cancers using DNNs and gene expression profiling as well[127]. LIME[128] is a popular sensitivity analysis method which learns an interpretable model locally around the prediction. Simonyan et al.[124] proposed using the gradient of the output with respect to pixels of an input image to compute a saliency map of the image. To avoid the saturation effect in perturbation-based and gradient-based approaches, DeepLIFT[129] decomposes the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. SHAP[130] unifies these approaches using a theoretically grounded method which assigns each feature an importance value for a particular prediction. Finally, GNNExplainer[131] is a new approach among a family of methods which provide interpretable explanations for predictions of GNN-based models on graph-based DL tasks. Given an instance, GNNExplainer identifies a compact subgraph structure and a small subset of node features that have a crucial role in GNN's prediction.

The efforts towards developing tools for explanation of DNNs are still in their infancy and are rapidly growing; challenges still abound towards a fully explainable systems in biology. The key problem is that the current general purpose methods to explain DL models are not sufficient especially in the clinical settings[132]. For the scientist and clinicians to trust these black box models they need to be able to explain themselves in a human-understandable fashion with quantifiable level of uncertainty, summarize the reasons for their behaviours, and suggest the additional steps (e.g., experiments, clinical studies, etc.) required to be able to reliably defend their decisions. We speculate that the new generation of explainable methods focus on helping these black-box models to transition from hypothesis generation machines into hypothesis testing ones which can communicate easier with medical practitioners.

Training efficiency. Despite the high accuracy of many DL approaches, their performance often comes at a high monetary and computational cost. For example, the monetary cost of consumed power and computation time is estimated to be up to hundreds of thousands of US dollars to train a single model[133]. The extreme costs of large DL models can prevent broader research community from reproducing and improving upon the current results. Thus, it is practical to consider lower-cost alternatives that are available and feasible for researchers with more modest resources. These issues are relevant for applying DL to computational biology. For instance, training the state-of-the-art

protein structure prediction model AlphaFold2 requires computational resources equivalent to 100–200 GPUs running for a few weeks[21]. In the following paragraphs, we discuss common strategies utilized by the DL community to decrease the memory and computation cost in training, and potential directions for applying similar strategies to improve the efficiency of DL models in computational biology.

The most direct method of reducing the training cost of a DL method is to perform transfer learning on the available pretrained general model, instead of training the new model from scratch. It's a common approach in training DL models for NLP tasks, and it has been shown that general language knowledge models are a good starting point for various different NLP tasks[134]. This approach can be adopted in computational biology, if all downstream tasks can start with a general model on biological data. For example, Zaheer et al.[135] trained a general human DNA sequence model based on human reference genome GRCh37, with self-supervised learning (masked DNA sequence prediction and next DNA sequence segment prediction). Subsequently, they have shown successful downstream task (Promoter Region Prediction) performance by solely applying transfer learning on the general model. Using pretrained models largely decreases (i) the size of task-specific datasets needed for training; and (ii) the total amount of local training needed for certain tasks that researchers are interested in. Thus creating general models that can be shared and used by the entire research community will greatly reduce the resources needed for training models on specific tasks by individual research groups. However, this approach will be less useful if the data distribution for different downstream tasks is drastically different compared with the data used by the general pretrained model. For instance, DeepVariant has limited applicability to non-human SNV calling due to the differences between diploid and haploid genomes, and nucleic acid distributions[4]. In these cases, we still need to train from scratch or spend a significant amount of resources on re-training the base model.

An alternative approach is to design DL model architectures with improved efficiency. As one of the most widely-studied architectures in DL, numerous low-cost variants of CNNs have been proposed. Some popular examples of efficient CNN architectures include the MobileNet family[136], DenseNet[137], EfficientNet[138], and CSPNet[139]. Similarly, numerous efficiency-based architectural modifications have been proposed for the transformer model, many of which aim to reduce the quadratic computational complexity incurred by the self-attention mechanism[140]. Additionally, some transformer architectural variants explore the use of parameter sharing and factorization to reduce the memory cost of model training[141]. Going further, efficient architectural variants have been discovered for RNNs[142] and graph neural networks (GNNs)[143,144], including specialized architectures that are tuned for better efficiency within the biological domain[145].

For computational biology applications, one approach for boosting efficiency relies on exploiting inherent sparsity and locality of biological data (e.g. focusing only on the SNV calls rather than the whole genome[146]). Researchers are also using transformers for DNA/RNA sequence modeling[135], but transformer models have high training costs due to the expensive global attention mechanism. Prior domain expertise can be leveraged here to help prune attention neighborhoods, and subsequently improve training efficiency of the models. Finally, one can also change the model's architecture during training to adaptively improve the training efficiency. The practice of model pruning, which removes unimportant parameters from the model, has become a popular method of deriving lightweight DL models[147] in deployment.

As the amount of biological data keeps increasing, the size of the neural networks will increase as well, and lead to a higher total number of training iterations required for convergence. Therefore it's natural to explore dataset reduction strategies as one of solutions to the efficiency challenge. One feasible proposal is to construct coresets of the training dataset[148]. This can be done by using clustering methods on the dataset and choosing centroids as the representatives of the dataset. Alternatively, dataset condensation can be achieved by selecting the data samples that can best approximate the effect of training the model on the whole dataset. An orthogonal way of solving the high training cost problem for DL is to distribute the training on several cheap low-end devices. This step will decrease the total training time by distributing training, and decrease the total budget by using multiple cheap devices with less computation power. In general, the major distributed training methods are data parallelism, model parallelism and hybrid parallel training. Data parallel training splits and distributes parts of the dataset to each device[149], where model parallel training splits and distributes parts of the model to each device[150]. As all above methods are task-agnostic, they can be readily applied to DL models for computational biology.

Concluding comments. In summary, while the success of DL in areas such as protein structure prediction is paradigm shifting, other areas such as function prediction, genome engineering, and multi-omics are also observing rapid gains in performance compared to traditional approaches. For other areas such as phylogenetics, classical computational approaches seem to have the upper hand in those areas. Additional advances specific to DL applied to challenges across the biosciences will further leverage domain-specific biological knowledge while striving for high explainability and improved efficiency.

Received: 25 August 2021; Accepted: 9 March 2022;
Published online: 01 April 2022

References

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
3. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
4. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36,** 983–987 (2018).
5. Arango-Argoty, G. et al. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 1–15 (2018).
6. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
7. Nielsen, A. A. & Voigt, C. A. Deep learning to predict the lab-of-origin of engineered DNA. *Nat. Commun.* **9**, 1–10 (2018).
8. Pearce, R. & Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **297,** 100870 (2021).
9. AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
10. de Chadarevian, S. John Kendrew and myoglobin: Protein structure determination in the 1950s. *Protein Sci.* **27**, 1136–1143 (2018).
11. Stollar, E. J. & Smith, D. P. Uncovering protein structure. *Essays Biochem.* **64**, 649–680 (2020).
12. Srivastava, A., Nagai, T., Srivastava, A., Miyashita, O. & Tama, F. Role of computational methods in going beyond X-ray crystallography to explore protein structure and dynamics. *Int. J. Mol. Sci.* **19**, 3401 (2018).
13. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biol.* **13**, e1005324 (2017).
14. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Struct. Funct. Bioinforma.* **87**, 1149–1164 (2019).

15. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science.* **373**, 871–876 (2021).
16. Mirabello, C. & Wallner, B. RAWMSA: End-to-end deep learning using raw multiple sequence alignments. *PloS One* **14**, e0220182 (2019).
17. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* https://doi.org/10.1038/s41586-021-03828-1 (2021).
18. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics.* **35**, 4862–4865 (2019).
19. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. Learning protein structure with a differentiable simulator. In *International Conference on Learning Representations* (2018).
20. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst.* **8**, 292–301 (2019).
21. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
22. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
23. Madeira, F. et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
24. Bonetta, R. & Valentino, G. Machine learning techniques for protein function prediction. *Proteins: Struct. Funct. Bioinforma.* **88**, 397–413 (2020).
25. Huntley, R. P. et al. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.* **43**, D1057–D1063 (2015).
26. Zhang, M.-L. & Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **26**, 1819–1837 (2013).
27. Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
28. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
29. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
30. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics.* **34**, 660–668 (2018).
31. Alshahrani, M. et al. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics.* **33**, 2723–2730 (2017).
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
33. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics.* **36**, 422–429 (2020).
34. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
35. Chicco, D., Sadowski, P. & Baldi, P. Deep autoencoder neural networks for gene ontology annotation predictions. In *Proceedings of the 5th ACM Conference On Bioinformatics, Computational Biology, and Health Informatics*, 533–540 (2014).
36. Miranda, L. J. & Hu, J. A deep learning approach based on stacked denoising autoencoders for protein function prediction. In 2018 *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, 480–485 (IEEE, 2018).
37. Gligorijević, V., Barot, M. & Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics.* **34**, 3873–3881 (2018).
38. Zou, Z., Tian, S., Gao, X. & Li, Y. mlDEEpre: multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Front. Genet.* **9**, 714 (2019).
39. Li, S. et al. Deep learning-based prediction of species-specific protein S-glutathionylation sites. *Biochim. Biophys. Acta Proteins Proteom.* **1868**, 140422 (2020).
40. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2019).
41. Zhang, F. et al. Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics.* **19**, 1900019 (2019).
42. Aghazadeh, A. et al. Epistatic net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nat. Commun.* **12**, 1–10 (2021).
43. Brookes, D. H., Aghazadeh, A. & Listgarten, J. On the sparsity of fitness functions and implications for learning. In *Proceedings of the National Academy of Sciences* 119 (2022). https://www.pnas.org/content/119/1/e2109649118.
44. van Overbeek, M. et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell.* **63**, 633–646 (2016).
45. Cui, Y., Xu, J., Cheng, M., Liao, X. & Peng, S. Review of CRISPR/Cas9 sgRNA design tools. *Interdiscip. Sci. Computational Life Sci.* **10**, 455–465 (2018).
46. Xue, L., Tang, B., Chen, W. & Luo, J. Prediction of CRISPR sgRNA activity using a deep convolutional neural network. *J. Chem. Inf. Modeling.* **59**, 615–624 (2018).
47. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2019).
48. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 1–18 (2018).
49. Liu, Q., Cheng, X., Liu, G., Li, B. & Liu, X. Deep learning improves the ability of sgRNA off-target propensity prediction. *BMC Bioinforma.* **21**, 1–15 (2020).
50. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 1–14 (2019).
51. Liu, Q., He, D. & Xie, L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas system using attention boosted deep learning and network-based gene feature. *PLoS Computational Biol.* **15**, e1007480 (2019).
52. Zhang, G., Zeng, T., Dai, Z. & Dai, X. Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. *Computational Struct. Biotechnol. J.* **19**, 1445–1457 (2021).
53. Metsky, H. C. et al. Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01213-5 (2022).
54. Leenay, R. T. et al. Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nat. Biotechnol.* **37**, 1034–1037 (2019).
55. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2019).
56. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature.* **563**, 646–651 (2018).
57. Li, V. R., Zhang, Z. & Troyanskaya, O. G. CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes. *Bioinformatics.* **37**, i342–i348 (2021).
58. Kim, H. K. et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239 (2018).
59. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat. Biotechnol.* **38**, 1037–1043 (2020).
60. Kim, H. K. et al. Predicting the efficiency of prime editing guide RNAs in human cells. *Nat. Biotechnol.* **39**, 198–206 (2021).
61. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
62. Aghazadeh, A., Ocal, O. & Ramchandran, K. CRISPRLand: Interpretable large-scale inference of DNA repair landscape based on a spectral approach. *Bioinformatics.* **36**, i560–i568 (2020).
63. Sun, B., Feng, J. & Saenko, K. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 30 (2016).
64. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
65. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8-D13 (2018).
66. Athar, A. et al. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).
67. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
68. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
69. Barretina, J. et al. The Cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
70. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
71. Consortium, E. P. et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science.* **306**, 636–640 (2004).
72. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* **8**, 118–127 (2007).
73. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
74. Leek, J. T. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161–e161 (2014).
75. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
76. Zhu, Q. et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods.* **12**, 211–214 (2015).
77. Wong, A. K., Krishnan, A. & Troyanskaya, O. G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res.* **46**, W65–W70 (2018).
78. Yao, V. et al. An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* **36**, 1091–1099 (2018).
79. Ellis, S. E., Collado-Torres, L., Jaffe, A. & Leek, J. T. Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Res.* **46**, e54 (2018).

80. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
81. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
82. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods.* **16**, 1289–1296 (2019).
83. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods.* **15**, 1053–1058 (2018).
84. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
85. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 1–14 (2019).
86. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 1–13 (2017).
87. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 1–10 (2019).
88. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).
89. Wang, H., Lengerich, B. J., Aragam, B. & Xing, E. P. Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics.* **35**, 1181–1187 (2019).
90. Li, Z., Chang, C., Kundu, S. & Long, Q. Bayesian generalized biclustering analysis via adaptive structured shrinkage. *Biostatistics.* **21**, 610–624 (2020).
91. Argelaguet, R. et al. Multi-omics factor analysis – a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
92. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
93. Tong, L., Mitchel, J., Chatlin, K. & Wang, M. D. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med. Inform. Decis. Mak.* **20**, 1–12 (2020).
94. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell.* **177**, 1888–1902 (2019).
95. Mariette, J. & Villa-Vialaneix, N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics.* **34**, 1009–1015 (2018).
96. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 1–19 (2017).
97. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad. Sci.* **118,** e2023070118 (2021).
98. Buggenthin, F. et al. Prospective identification of hematopoietic lineage choice by deep learning. *Nat. Methods.* **14**, 403–406 (2017).
99. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinforma.* **35**, i501–i509 (2019).
100. Ma, T. & Zhang, A. multi-view factorization autoencoder with network constraints for multi-omic integrative analysis. In 2018 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 702-707 (IEEE, 2018).
101. Lee, G., Nho, K., Kang, B., Sohn, K.-A. & Kim, D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* **9**, 1952 (2019).
102. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods.* **11**, 333–337 (2014).
103. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.* **20**, 1–17 (2019).
104. Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**, 605–615 (2007).
105. Beery, A. K. & Zucker, I. Sex bias in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* **35**, 565–572 (2011).
106. Carlson, C. S. et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661 (2013).
107. Felsenstein, J. *Inferring Phylogenies*, vol. 2 (Sinauer Associates Sunderland, MA, 2004).
108. Nute, M., Saleh, E. & Warnow, T. Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. *Syst. Biol.* **68**, 396–411 (2018).
109. Nakhleh, L. In *Problem solving handbook in computational biology and bioinformatics*, 125–158 (Springer, 2010).
110. Dopazo, J. & Carazo, J. M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. evolution.* **44**, 226–233 (1997).
111. Suvorov, A., Hochuli, J. & Schrider, D. R. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Syst. Biol.* **69**, 221–233 (2020).
112. Zou, Z., Zhang, H., Guan, Y. & Zhang, J. Deep residual neural networks resolve quartet molecular phylogenies. *Mol. Biol. Evolution* **37**, 1495–1507 (2020).
113. Zaharias, P., Grosshauser, M. & Warnow, T. Re-evaluating deep neural networks for phylogeny estimation: the issue of taxon sampling *J Comput Biol* **29**, 74-89 (2021).
114. Jiang, Y., Balaban, M., Zhu, Q. & Mirarab, S. DEPP: Deep learning enables extending species trees using single genes. https://doi.org/10.1101/2021.01.22.427808(2021).
115. Bhattacharjee, A. & Bayzid, M. S. Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. *BMC genomics.* **21**, 1–14 (2020).
116. Linial, N., London, E. & Rabinovich, Y. The geometry of graphs and some of its algorithmic applications. *Combinatorica.* **15**, 215–245 (1995).
117. Nickel, M. & Kiela, D. Poincaré embeddings for learning hierarchical representations. *Adv. Neural Inf. Process. Syst.* **30**, 6338–6347 (2017).
118. Abadi, S., Avram, O., Rosset, S., Pupko, T. & Mayrose, I. ModelTeller: model selection for optimal phylogenetic reconstruction using machine learning. *Mol. Biol. Evolution* **37**, 3338–3352 (2020).
119. Kumar, S. & Sharma, S. Evolutionary sparse learning for phylogenomics. *Mol. Biol. Evolution.* **38**, 4674–4682 (2021).
120. Azer, E. S., Ebrahimabadi, M. H., Malikić, S., Khardon, R. & Sahinalp, S. C. Tumor phylogeny topology inference via deep learning. *iScience.* **23**, 101655 (2020).
121. Bello, I., Pham, H., Le, Q. V., Norouzi, M. & Bengio, S. Neural combinatorial optimization with reinforcement learning. In *Workshop at International Conference on Learning Representations, ICLR'17.* (2017).
122. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process.* **73**, 1–15 (2018).
123. Berkes, P. & Wiskott, L. On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural Comput.* **18**, 1868–1895 (2006).
124. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations* (2014).
125. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* **10**, e0130140 (2015).
126. Zurada, J. M., Malinowski, A. & Cloete, I. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94*, vol. 6, 447–450 (IEEE, 1994).
127. Khan, J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**, 673–679 (2001).
128. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (2016).
129. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning,* 3145–3153 (PMLR, 2017).
130. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (2017).
131. Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. GNNExplainer: Generating explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* **32**, 9240 (2019).
132. Gilpin, L. H. et al. Explaining explanations: An overview of interpretability of machine learning. In 2018 *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89 (IEEE, 2018).
133. Yang, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **32** (2019).
134. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171-4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). https://aclanthology.org/N19-1423.
135. Zaheer, M. et al. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems* (*NeurIPS*), **33**, 17283–17297 (2020).

136. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520 (2018).
137. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708 (2017).
138. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105-6114 (PMLR, 2019).
139. Wang, C.-Y. et al. Cspnet: A new backbone that can enhance learning capability of CNN. In Proceedings of *the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 390–391 (2020).
140. Wu, Z., Liu, Z., Lin, J., Lin, Y. & Han, S. Lite transformer with long-short range attention. In *International Conference on Learning Representations* (2019).
141. Lan, Z. et al. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th *International Conference on Learning Representations, ICLR 2020,* Addis Ababa, Ethiopia, April 26–30, 2020 (2020).
142. Kusupati, A. et al. Fastgrnn: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. *Adv. Neural Inf. Process. Syst.* **31**, 9031–9042 (2018).
143. Chiang, W.-L. et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 257–266 (2019).
144. Zou, D. et al. Layer-dependent importance sampling for training deep and large graph convolutional networks. *Adv. Neural Inf. Process. Syst.* **32**, 11249–11259 (2019).
145. Pouladi, F., Salehinejad, H. & Gilani, A. M. Recurrent neural networks for sequential phenotype prediction in genomics. In 2015 *International Conference on Developments of E-Systems Engineering (DeSE)*, 225–230 (IEEE, 2015).
146. Ke, Z. & Vikalo, H. A convolutional auto-encoder for haplotype assembly and viral quasispecies reconstruction. *Adv. Neural Inf. Process. Syst.* **33**, 13493–13503 (2020).
147. Liu, Z., Sun, M., Zhou, T., Huang, G. & Darrell, T. Rethinking the value of network pruning. In *International Conference on Learning Representations* (2018).
148. Mirzasoleiman, B., Bilmes, J. & Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 6950–6960 (PMLR, 2020).
149. Lin, T., Stich, S. U., Patel, K. K. & Jaggi, M. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations* (2019).
150. Geng, J., Li, D. & Wang, S. Elasticpipe: An efficient and dynamic model-parallel solution to DNN training. In *Proceedings of the 10th Workshop on Scientific Cloud Computing*, 5–9 (2019).

Acknowledgements

A.A. is supported by the ARO (W911NF2110117). R.B. and CJ.B. are supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571, N00014-20-1-2534, and MURI N00014-20-1-2787; AFOSR grant FA9550-18-1-0478; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. M.N and R.A.L.E. are supported by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science (T15LM007093). D.A.A. is partially supported by funds from the University of Houston. A.B., B.K., N.S., and T.J.T are partially supported by funds from the FunGCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. T.J.T is supported by NIH grant P01AI152999 and by NSF grant EF-212638. B.K. is supported by a fellowship from the National Library of Medicine Training Program in Biomedical Informatics and Data Science (5T15LM007093-30, PI: Kavraki). Z.Y., M.E., and L.N. are supported by NSF grants DBI-2030604 and IIS-2106837. R.D. and V.Y. are supported by Cancer Prevention & Research Institute of Texas (CPRIT) Award (RR190065). V.Y. is a CPRIT Scholar in Cancer Research and also supported by NIH grant RF1AG054564. A.K. is supported by NSF grants CCF-1907936, CNS-2003137.

Author contributions

N.S. and A.A. designed figures and conceptualized the manuscript. N.S., A.A., R.A.L.E, and B.K. contributed text to the introduction and general challenges for deep learning in the biosciences sections. A.A. contributed text for the genome engineering section. D.A.A. contributed text for the protein structure prediction section. A.B. contributed text for the protein function prediction section. R.B., C.J.B., and A.A. contributed text for the explainability section. C.D., C.R.W., and A.K. contributed text for the training efficiency section. R.D. and V.Y. contributed text for the systems biology and data integration section. M.E., M.G.N., L.N., and Z.Y. contributed text and figures for the phylogenetics section. T.J.T. supervised the work and contributed to manuscript conceptualization. All authors have edited and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Todd J. Treangen.

Peer review information *Nature Communications* thanks Bharath Ramsundar, Aurelien Tellier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

Chicco and Jurman *BMC Genomics* (2020) 21:6
https://doi.org/10.1186/s12864-019-6413-7

BMC Genomics

RESEARCH ARTICLE Open Access

The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation

Davide Chicco[1,2*] and Giuseppe Jurman[3]

Abstract

Background: To evaluate binary classifications and their confusion matrices, scientific researchers can employ several statistical rates, accordingly to the goal of the experiment they are investigating. Despite being a crucial issue in machine learning, no widespread consensus has been reached on a unified elective chosen measure yet. Accuracy and F_1 score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification tasks. However, these statistical measures can dangerously show overoptimistic inflated results, especially on imbalanced datasets.

Results: The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

Conclusions: In this article, we show how MCC produces a more informative and truthful score in evaluating binary classifications than accuracy and F_1 score, by first explaining the mathematical properties, and then the asset of MCC in six synthetic use cases and in a real genomics scenario. We believe that the Matthews correlation coefficient should be preferred to accuracy and F_1 score in evaluating binary classification tasks by all scientific communities.

Keywords: Matthews correlation coefficient, Binary classification, F_1 score, Confusion matrices, Machine learning, Biostatistics, Accuracy, Dataset imbalance, Genomics

Background

Given a clinical feature dataset of patients with cancer traits [1, 2], which patients will develop the tumor, and which will not? Considering the gene expression of neuroblastoma patients [3], can we identify which patients are going to survive, and which will not? Evaluating the metagenomic profiles of patients [4], is it possible to discriminate different phenotypes of a complex disease? Answering these questions is the aim of machine learning and computational statistics, nowadays pervasive in analysis of biological and health care datasets, and many other scientific fields. In particular, these binary classification tasks can be efficiently addressed by supervised machine learning techniques, such as artificial neural networks [5], *k*-nearest neighbors [6], support vector machines [7], random forest [8], gradient boosting [9], or other methods. Here the word *binary* means that the data element statuses and prediction outcomes (class labels) can be twofold: in the example of patients, it can mean healthy/sick, or low/high grade tumor. Usually scientists indicate the two classes as the negative and the positive class. The term *classification* means that the goal of the process is to attribute the correct label to each data instance (sample); the process itself is known as the classifier, or classification algorithm.

*Correspondence: davidechicco@davidechicco.it
[1]Krembil Research Institute, Toronto, Ontario, Canada
[2]Peter Munk Cardiac Centre, Toronto, Ontario, Canada
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Scientists have used binary classification to address several questions in genomics in the past, too. Typical cases include the application of machine learning methods to microarray gene expressions [10] or to single-nucleotide polymorphisms (SNPs) [11] to classify particular conditions of patients. Binary classification can also be used to infer knowledge about biology: for example, computational intelligence applications to ChIP-seq can predict transcription factors [12], applications to epigenomics data can predict enhancer-promoter interactions [13], and applications to microRNA can predict genomic inverted repeats (pseudo-hairpins) [14].

A crucial issue naturally arises, concerning the outcome of a classification process: how to evaluate the classifier performance? A relevant corpus of published works has stemmed until today throughout the last decades for possible alternative answers to this inquiry, by either proposing a novel measure or comparing a subset of existing ones on a suite of benchmark tasks to highlight pros and cons [15–28], also providing off-the-shelf software packages [29, 30]. Despite the amount of literature dealing with this problem, this question is still an open issue. However, there are several consolidated and well known facts driving the choice of evaluating measures in the current practice.

Accuracy, MCC, F_1 score. Many researchers think the most reasonable performance metric is the ratio between the number of correctly classified samples and the overall number of samples (for example, [31]). This measure is called *accuracy* and, by definition, it also works when labels are more than two (multiclass case). However, when the dataset is unbalanced (the number of samples in one class is much larger than the number of samples in the other classes), accuracy cannot be considered a reliable measure anymore, because it provides an overoptimistic estimation of the classifier ability on the majority class [32–35].

An effective solution overcoming the class imbalance issue comes from the *Matthews correlation coefficient (MCC)*, a special case of the ϕ phi coefficient [36].

Stemming from the definition of the phi coefficient, a number of metrics have been defined and mainly used for purposes other than classification, for instance as association measures between (discrete) variables, with the *Cramér's V* (or *Cramér's* ϕ) being one of the most common rates [37].

Originally developed by Matthews in 1975 for comparison of chemical structures [38], MCC was re-proposed by Baldi and colleagues [39] in 2000 as a standard performance metric for machine learning with a natural extension to the multiclass case [40]. MCC soon started imposing as a successful indicator: for instance, the Food and Drug Administration (FDA) agency of the USA employed the MCC as the main evaluation measure in the MicroArray II / Sequencing Quality Control (MAQC/SEQC) projects [41, 42]. The effectiveness of MCC has been shown in other scientific fields as well [43, 44].

Although being widely acknowledged as a reliable metric, there are situations - albeit extreme - where either MCC cannot be defined or it displays large fluctuations [45], due to imbalanced outcomes in the classification. Even if mathematical workarounds and Bayes-based improvements [46] are available for these cases, they have not been adopted widely yet.

Shifting context from machine learning to information retrieval, and thus interpreting positive and negative class as relevant and irrelevant samples respectively, the recall (that is the accuracy on the positive class) can be seen as the fraction of relevant samples that are correctly retrieved. Then its dual metric, the precision, can be defined as the fraction of retrieved documents that are relevant. In the learning setup, the pair precision/recall provides useful insights on the classifier's behaviour [47], and can be more informative than the pair specificity/sensitivity [48]. Meaningfully combining precision and recall generates alternative performance evaluation measures. In particular, their harmonic mean has been originally introduced in statistical ecology by Dice [49] and Sørensen [50] independently in 1948, then rediscovered in the 1970s in information theory by van Rijsbergen [51, 52] and finally adopting the current notation of F_1 measure in 1992 [53]. In the 1990s, in fact, F_1 gained popularity in the machine learning community, to the point that it was also re-introduced later in the literature as a novel measure [54].

Nowadays, the F_1 measure is widely used in most application areas of machine learning, not only in the binary scenario, but also in multiclass cases. In multiclass cases, researchers can employ the F_1 micro/macro averaging procedure [55–60], which can be even targeted for *ad-hoc* optimization [61].

The distinctive features of F_1 score have been discussed in the literature [62–64]. Two main properties characterize F_1 from MCC. First, F_1 varies for class swapping, while MCC is invariant if the positive class is renamed negative and vice versa. This issue can be overcome by extending the macro/micro averaging procedure to the binary case itself [17], by defining the F_1 score both on the positive and negative classes and then average the two values (macro), and using the average sensitivity and average precision values (micro). The micro/macro averaged F_1 is invariant for class swapping and its behaviour is more similar to MCC. However, this procedure is biased [65], and it is still far from being accepted as a standard practice by the community. Second, F_1 score is independent from the number of samples correctly classified as negative. Recently, several scientists highlighted drawbacks of the

F_1 measure [66, 67]: in fact, Hand and Peter [68] claim that alternative measures should be used instead, due to its major conceptual flaws. Despite the criticism, F_1 remains one of the most widespread metrics among researchers. For example, when Whalen and colleagues released *TargetFinder*, a tool to predict enhancer-promoters interactions in genomics, they showed its results measured only by F_1 score [13], making it impossible to detect the actual true positive rate and true negative rate of their tests [69].

Alternative metrics. The current most popular and widespread metrics include *Cohen's kappa* [70–72]: originally developed to test inter-rater reliability, in the last decades Cohen's kappa entered the machine learning community for comparing classifiers' performances. Despite its popularity, in the learning context there are a number of issues causing the kappa measure to produce unreliable results (for instance, its high sensitivity to the distribution of the marginal totals [73–75]), stimulating research for more reliable alternatives [76]. Due to these issues, we chose not to include Cohen's kappa in the present comparison study.

In the 2010s, several alternative novel measures have been proposed, either to tackle a particular issue such as imbalance [34, 77], or with a broader purpose. Among them, we mention the *confusion entropy* [78, 79], a statistical score comparable with MCC [80], and the *K measure* [81], a theoretically grounded measure that relies on a strong axiomatic base.

In the same period, Powers proposed *informedness* and *markedness* to evaluate binary classification confusion matrices [22]. Powers defines informedness as *true positive rate – true negative rate*, to express how the predictor is informed in relation to the opposite condition [22]. And Powers defines markedness as *precision – negative predictive value*, meaning the probability that the predictor correctly marks a specific condition [22].

Other previously introduced rates for confusion matrix evaluations are *macro average arithmetic* (MAvA) [18], *geometric mean* (Gmean or G-mean) [82], and *balanced accuracy* [83], which all represent classwise weighted accuracy rates.

Notwithstanding their effectiveness, all the aforementioned measures have not yet achieved such a diffusion level in the literature to be considered solid alternatives to MCC and F_1 score. Regarding MCC and F_1, in fact, Dubey and Tatar [84] state that these two measure "provide more realistic estimates of real-world model performance".

However, there are many instances where MCC and F_1 score disagree, making it difficult for researchers to draw correct deductions on the behaviour of the investigated classifier.

MCC, F_1 score, and accuracy can be computed when a specific statistical threshold τ for the confusion matrix is set. When the confusion matrix threshold is not unique, researchers can instead take advantage of classwise rates: *true positive rate* (or *sensitivity*, or *recall*) and *true negative rate* (or *specificity*), for example, computed for all the possible confusion matrix thresholds. Different combinations of these two metrics give rise to alternative measures: among them, the area under the *receiver operating characteristic curve* (AUROC or ROC AUC) [85–91] plays a major role, being a popular performance measure when a singular threshold for the confusion matrix is unavailable. However, ROC AUC presents several flaws [92], and it is sensitive to class imbalance [93]. Hand and colleagues proposed improvements to address these issues [94], that were partially rebutted by Ferri and colleagues [95] some years later.

Similar to ROC curve, the *precision-recall (PR) curve* can be used to test all the possible positive predictive values and sensitivities obtained through a binary classification [96]. Even if less common than the ROC curve, several scientists consider the PR curve more informative than the ROC curve, especially on imbalanced biological and medical datasets [48, 97, 98].

If no confusion matrix threshold is applicable, we suggest the readers to evaluate their binary evaluations by checking both the PR AUC and the ROC AUC, focusing on the former [48, 97]. If a confusion matrix threshold is at disposal, instead, we recommend the usage of the Matthews correlation coefficient over F_1 score, and accuracy.

In this manuscript, we outline the advantages of the Matthews correlation coefficient by first describing its mathematical foundations and its competitors accuracy and F_1 score ("Notation and mathematical foundations" section), and by exploring their relationships afterwards (Relationships between rates). We decided to focus on accuracy and F_1 score because they are the most common metrics used for binary classification in machine learning. We then show some examples to illustrate why the MCC is more robust and reliable than F_1 score, on six synthetic scenarios ("Use cases" section) and a real genomics application ("Genomics scenario: colon cancer gene expression" section). Finally, we conclude the manuscript with some take-home messages ("Conclusions" section).

Methods

Notation and mathematical foundations

Setup. The framework where we set our investigation is a machine learning task requiring the solution of binary classification problem. The dataset describing the task is composed by n^+ examples in one class, labeled *positive*, and n^- examples in the other class, called *negative*. For instance, in a biomedical case control study, the healthy individuals are usually labelled negative, while the positive label is usually attributed to the sick patients. As a

general practice, given two phenotypes, the positive class corresponds to the *abnormal* phenotype. This ranking is meaningful for example, in different stages of a tumor.

The classification model forecasts the class of each data instance, attributing to each sample its predicted label (positive or negative): thus, at the end of the classification procedure, every sample falls in one of the following four cases:

- Actual positives that are correctly predicted positives are called *true positives* (TP);
- Actual positives that are wrongly predicted negatives are called *false negatives* (FN);
- Actual negatives that are correctly predicted negatives are called *true negatives* (TN);
- Actual negatives that are wrongly predicted positives are called *false positives* (FP).

This partition can be presented in a 2×2 table called *confusion matrix* $\mathbf{M} = \begin{pmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{pmatrix}$ (expanded in Table 1), which completely describes the outcome of the classification task.

Clearly $\mathrm{TP} + \mathrm{FN} = n^+$ and $\mathrm{TN} + \mathrm{FP} = n^-$. When one performs a machine learning binary classification, she/he hopes to see a high number of true positives (TP) and true negatives (TN), and less false negatives (FN) and false positives (FP). When $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ 0 & n^- \end{pmatrix}$ the classification is perfect.

Since analyzing all the four categories of the confusion matrix separately would be time-consuming, statisticians introduced some useful statistical rates able to immediately describe the quality of a prediction [22], aimed at conveying into a single figure the structure of **M**. A set of these functions act classwise (either actual or predicted), that is, they involve only the two entries of **M** belonging to the same row or column (Table 2). We cannot consider such measures fully informative because they use only two categories of the confusion matrix [39].

Accuracy. Moving to global metrics having three or more entries of **M** as input, many researchers consider computing the accuracy as the standard way to go. Accuracy, in fact, represents the ratio between the correctly predicted instances and all the instances in the dataset:

$$\text{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{n^+ + n^-} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \qquad (1)$$

Table 1 The standard confusion matrix M

	Predicted positive	Predicted negative
Actual positive	True positives TP	False negatives FN
Actual negative	False positives FP	True negatives TN

True positives (TP) and true negatives (TN) are the correct predictions, while false negatives (FN) and false positives (FP) are the incorrect predictions

Table 2 Classwise performance measures

Sensitivity, recall, true positive rate	$= \frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{\mathrm{TP}}{n^+}$	Specificity, true negative rate	$= \frac{\mathrm{TN}}{\mathrm{TN+FP}} = \frac{\mathrm{TN}}{n^-}$
Positive predictive value, precision	$= \frac{\mathrm{TP}}{\mathrm{TP+FP}}$	Negative predictive value	$= \frac{\mathrm{TN}}{\mathrm{TN+FN}}$
False positive rate, fallout	$= \frac{\mathrm{FP}}{\mathrm{FP+TN}} = \frac{\mathrm{FP}}{n^-}$	False discovery rate	$= \frac{\mathrm{FP}}{\mathrm{FP+TP}}$

TP: true positives. TN: true negatives. FP: false positives. FN: false negatives

(worst value: 0; best value: 1)

By definition, the accuracy is defined for every confusion matrix **M** and ranges in the real unit interval $[0, 1]$; the best value 1.00 corresponds to perfect classification $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ 0 & n^- \end{pmatrix}$ and the worst value 0.00 corresponds to perfect misclassification $\mathbf{M} = \begin{pmatrix} 0 & n^+ \\ n^- & 0 \end{pmatrix}$.

As anticipated (Background), accuracy fails in providing a fair estimate of the classifier performance in the class-unbalanced datasets. For any dataset, the proportion of samples belonging to the largest class is called the *no-information error rate* $\mathrm{ni} = \frac{\max\{n^+, n^-\}}{n^+ + n^-}$; a binary dataset is (perfectly) balanced if the two classes have the same size, that is, $\mathrm{ni} = \frac{1}{2}$, and it is unbalanced if one class is much larger than the other, that is $\mathrm{ni} \gg \frac{1}{2}$. Suppose now that $\mathrm{ni} \neq \frac{1}{2}$, and apply the trivial majority classifier: this algorithm learns only which is the largest class in the training set, and attributes this label to all instances. If the largest class is the positive class, the resulting confusion matrix is $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ n^- & 0 \end{pmatrix}$, and thus $\text{accuracy} = \mathrm{ni}$. If the dataset is highly unbalanced, $\mathrm{ni} \approx 1$, and thus the accuracy measure gives an unreliable estimation of the goodness of the classifier. Note that, although we achieved this result by mean of the trivial classifier, this is quite a common effect: as stated by Blagus and Lusa [99], several classifiers are biased towards the largest class in unbalanced studies.

Finally, consider another trivial algorithm, the coin tossing classifier: this classifier randomly attributes to each sample, the label positive or negative with probability $\frac{1}{2}$. Applying the coin tossing classifier to any binary dataset gives an accuracy with expected value $\frac{1}{2}$, since $\langle \mathbf{M} \rangle = \begin{pmatrix} n^+/2 & n^+/2 \\ n^-/2 & n^-/2 \end{pmatrix}$.

Matthews correlation coefficient (MCC). As an alternative measure unaffected by the unbalanced datasets issue, the Matthews correlation coefficient is a contingency matrix method of calculating the *Pearson product-moment correlation coefficient* [22] between actual and predicted values. In terms of the entries of **M**, MCC reads as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \cdot (\mathrm{TP}+\mathrm{FN}) \cdot (\mathrm{TN}+\mathrm{FP}) \cdot (\mathrm{TN}+\mathrm{FN})}} \quad (2)$$

(worst value: –1; best value: +1)

MCC is the only binary classification rate that generates a high score only if the binary predictor was able to correctly predict the majority of positive data instances and the majority of negative data instances [80, 97].

It ranges in the interval $[-1, +1]$, with extreme values –1 and +1 reached in case of perfect misclassification and perfect classification, respectively, while $\mathrm{MCC} = 0$ is the expected value for the coin tossing classifier.

A potential problem with MCC lies in the fact that MCC is undefined when a whole row or column of **M** is zero, as it happens in the previously cited case of the trivial majority classifier. However, some mathematical considerations can help meaningfully fill in the gaps for these cases. If **M** has only one non-zero entry, this means that all samples in the dataset belong to one class, and they are either all correctly (for $\mathrm{TP} \neq 0$ or $\mathrm{TN} \neq 0$) or incorrectly (for $\mathrm{FP} \neq 0$ or $\mathrm{FN} \neq 0$) classified. In this situations, $\mathrm{MCC} = 1$ for the former case and $\mathrm{MCC} = -1$ for the latter case. We are then left with the four cases where a row or a column of **M** are zero, while the other two entries are non zero. That is, when **M** is one of $\begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix}$, $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 \\ b & a \end{pmatrix}$ or $\begin{pmatrix} 0 & b \\ 0 & a \end{pmatrix}$, with $a, b \geq 1$: n in all four cases, MCC takes the indefinite form $\frac{0}{0}$. To detect a meaningful value of MCC for these four cases, we proceed through a simple approximation via a calculus technique. If we substitute the zero entries in the above matrices with the arbitrarily small value ϵ, in all four cases, we obtain

$$\begin{aligned} \mathrm{MCC} &= \frac{a\epsilon - b\epsilon}{\sqrt{(a+b)(a+\epsilon)(b+\epsilon)(\epsilon+\epsilon)}} \\ &= \frac{\epsilon}{\sqrt{\epsilon}} \frac{a-b}{\sqrt{2(a+b)(a+\epsilon)(b+\epsilon)}} \\ &\approx \sqrt{\epsilon} \frac{a-b}{\sqrt{2ab(a-b)}} \to 0 \quad \text{for } \epsilon \to 0 \end{aligned}$$

With these positions MCC is now defined for all confusion matrices **M**. As a consequences, $\mathrm{MCC} = 0$ for the trivial majority classifier, and 0 is also the expected value for the coin tossing classifier.

Finally, in some cases it might be useful to consider the *normalized* MCC, defined as $\mathrm{nMCC} = \frac{\mathrm{MCC}+1}{2}$, and linearly projecting the original range into the interval [0,1], with $\mathrm{nMCC} = \frac{1}{2}$ as the average value for the coin tossing classifier.

$\mathbf{F_1}$ score. This metric is the most used member of the parametric family of the F-measures, named after the parameter value $\beta = 1$. F_1 score is defined as the harmonic mean of precision and recall (Table 2) and as a function of **M**, has the following shape:

$$F_1 \text{ score} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (3)$$

(worst value: 0; best value: 1)

F_1 ranges in $[0, 1]$, where the minimum is reached for $\mathrm{TP} = 0$, that is, when all the positive samples are misclassified, and the maximum for $\mathrm{FN} = \mathrm{FP} = 0$, that is for perfect classification. Two main features differentiate F_1 from MCC and accuracy: F_1 is independent from TN, and it is not symmetric for class swapping.

F_1 is not defined for confusion matrices $\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & n^- \end{pmatrix}$: we can set $F_1 = 1$ for these cases. It is also worth mentioning that, when defining the F_1 score as the harmonic mean of precision and recall, the cases $\mathrm{TP} = 0$, $\mathrm{FP} > 0$, and $\mathrm{FN} > 0$ remain undefined, but using the expression $\frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$, the F_1 score is defined even for these confusion matrices and its value is zero.

When a trivial majority classifier is used, due to the asymmetry of the measure, there are two different cases: if $n^+ > n^-$, then $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ n^- & 0 \end{pmatrix}$ and $F_1 = \frac{2n^+}{2n^+ n^-}$, while if $n^- > n^+$ then $\mathbf{M} = \begin{pmatrix} 0 & n^+ \\ 0 & n^- \end{pmatrix}$, so that $F_1 = 0$. Further, for the coin tossing algorithm, the expected value is $F_1 = \frac{2n^+}{3n^+ + n^-}$.

Relationship between measures

After having introduced the statistical background of Matthews correlation coefficient and the other two measures to which we compare it (accuracy and F_1 score), we explore here the correlation between these three rates. To explore these statistical correlations, we take advantage of the Pearson correlation coefficient (PCC) [100], which is a rate particularly suitable to evaluate the linear relationship between two continuous variables [101]. We avoid the usage of rank correlation coefficients (such as Spearman's ρ and Kendall's τ [102]) because we are not focusing on the ranks for the two lists.

For a given positive integer $N \geq 10$, we consider all the possible $\binom{N+3}{3}$ confusion matrices for a dataset with N samples and, for each matrix, compute the accuracy, MCC and F_1 score and then the Pearson correlation coefficient for the three set of values. MCC and accuracy resulted strongly correlated, while the Pearson coefficient is less than 0.8 for the correlation of F_1 with the other two measures (Table 3). Interestingly, the correlation grows with N, but the increments are limited.

Similar to what Flach and colleagues did for their isometrics strategy [66], we depict a scatterplot of the MCCs and F_1 scores for all the 21 084 251 possible confusion matrices for a toy dataset with 500 samples (Fig. 1). We

Table 3 Correlation between MCC, accuracy, and F_1 score values

N	PCC (MCC, F_1 score)	PCC (MCC, accuracy)	PCC (accuracy, F_1 score)
10	0.742162	0.869778	0.744323
25	0.757044	0.893572	0.760708
50	0.766501	0.907654	0.769752
75	0.769883	0.912530	0.772917
100	0.771571	0.914926	0.774495
200	0.774060	0.918401	0.776830
300	0.774870	0.919515	0.777595
400	0.775270	0.920063	0.777976
500	0.775509	0.920388	0.778201
1 000	0.775982	0.921030	0.778652

Pearson correlation coefficient (PCC) between accuracy, MCC and F_1 score computed on all confusion matrices with given number of samples N

take advantage of this scatterplot to overview the mutual relations between MCC and F_1 score.

The two measures are reasonably concordant, but the scatterplot cloud is wide, implying that for each value of F_1 score there is a corresponding range of values of MCC and vice versa, although with different width. In fact, for any value $F_1 = \phi$, the MCC varies approximately between $[\phi - 1, \phi]$, so that the width of the variability range is 1, independent from the value of ϕ. On the other hand, for a given value MCC $= \mu$, the F_1 score can range in $[0, \mu + 1]$ if $\mu \leq 0$ and in $[\mu, 1]$ if $\mu > 0$, so that the width of the range is $1 - |\mu|$, that is, it depends on the MCC value μ.

Note that a large portion of the above variability is due to the fact that F_1 is independent from TN: in general, all matrices $\mathbf{M} = \begin{pmatrix} \alpha & \beta \\ \gamma & x \end{pmatrix}$ have the same value $F_1 = \frac{2\alpha}{2\alpha+\beta+\gamma}$ regardless of the value of x, while the corresponding MCC values range from $-\sqrt{\frac{\beta\gamma}{(\alpha+\beta)(\alpha+\gamma)}}$ for $x = 0$ to the asymptotic $\frac{a}{\sqrt{(\alpha+\beta)(\alpha+\gamma)}}$ for $x \to \infty$. For example, if we consider only the 63 001 confusion matrices of datasets of size 500 where TP=TN, the Pearson correlation coefficient between F_1 and MCC increases to 0.9542254.

Overall, accuracy, F_1, and MCC show reliable concordant scores for predictions that correctly classify both positives and negatives (having therefore many TP and TN), and for predictions that incorrectly classify both positives and negatives (having therefore few TP and TN); however, these measures show discordant behaviors when the prediction performs well just with one of the two binary classes. In fact, when a prediction displays many true positives but few true negatives (or many true negatives but few true positives) we will show that F_1 and accuracy can provide misleading information, while MCC

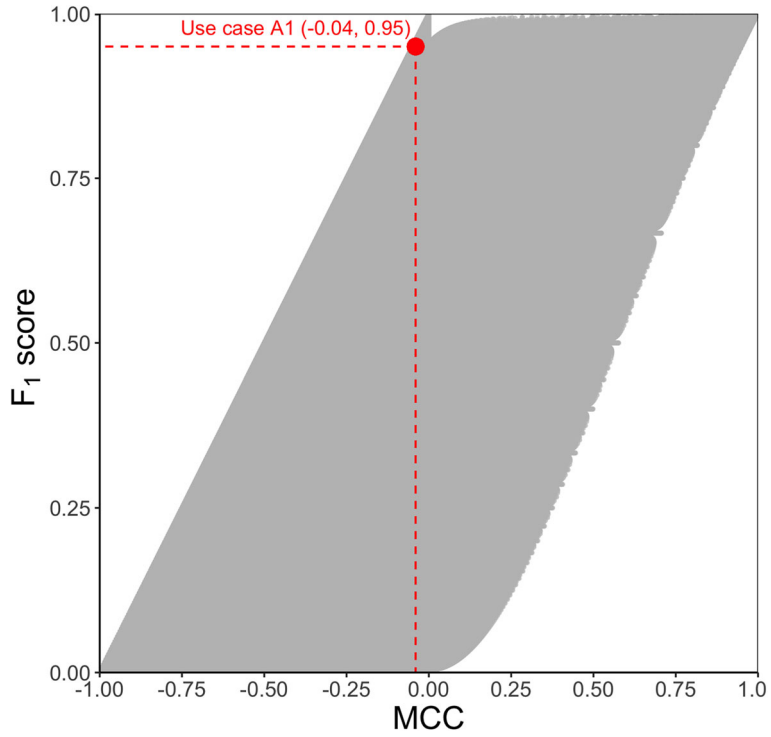


Fig. 1 Relationship between MCC and F_1 score. Scatterplot of all the 21 084 251 possible confusion matrices for a dataset with 500 samples on the MCC/F_1 plane. In red, the (−0.04, 0.95) point corresponding to use case A1

always generates results that reflect the overall prediction issues.

Results and discussion

Use cases

After having introduced the mathematical foundations of MCC, accuracy, and F_1 score, and having explored their relationships, here we describe some synthetic, realistic scenarios where MCC results are more informative and truthful than the other two measures analyzed.

Positively imbalanced dataset — Use case A1. Consider, for a clinical example, a positively imbalanced dataset made of 9 healthy individuals (negatives = 9%) and 91 sick patients (positives = 91%) (Fig. 2c). Suppose the machine learning classifier generated the following confusion matrix: TP=90, FN=1, TN=0, FP=9 (Fig. 2b).

In this case, the algorithm showed its ability to predict the positive data instances (90 sick patients out of 91 were correctly predicted), but it also displayed its lack of talent in identifying healthy controls (only 1 healthy individual out of 9 was correctly recognized) (Fig. 2b). Therefore, the overall performance should be judged poor. However, accuracy and of F_1 showed high values in this case: $accuracy = 0.90$ and $F_1\ score = 0.95$, both close to the best possible value 1.00 in the [0, 1] interval (Fig. 2a). At this point, if one decided to evaluate the performance of this classifier by considering only accuracy and F_1 score, he/she would overoptimistically think that the computational method generated excellent predictions.

Instead, if one decided to take advantage of the Matthews correlation coefficient in the Use case A1, he/she would notice the resulting MCC = −0.03 (Fig. 2a). By seeing a value close to zero in the [−1, +1] interval, he/she would be able to understand that the machine learning method has performed poorly.

Positively imbalanced dataset — Use case A2. Suppose the prediction generated this other confusion matrix: TP = 5, FN = 70, TN = 19, FP = 6 (Additional file 1b).

Here the classifier was able to correctly predict negatives (19 healthy individuals out of 25), but was unable to correctly identify positives (only 5 sick patients out of 70). In this case, all three statistical rates showed a low score

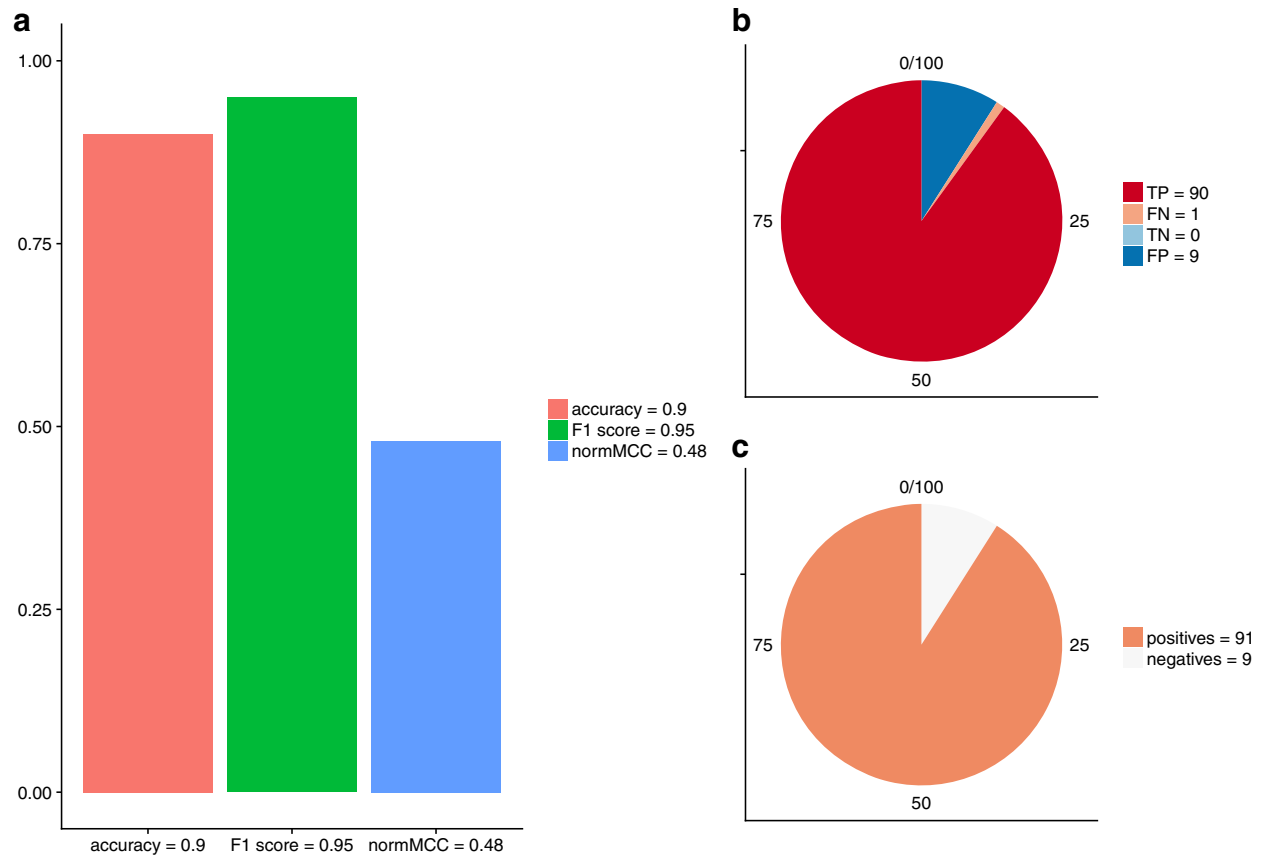


Fig. 2 Use case A1 — Positively imbalanced dataset. **a** Barplot representing accuracy, F_1, and normalized Matthews correlation coefficient ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case A1 positively imbalanced dataset. **b** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **c** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances

which emphasized the deficiency in the prediction process (*accuracy* = 0.24, F_1 *score* = 0.12, and *MCC* = −0.24).

Balanced dataset — Use case B1. Consider now, as another example, a balanced dataset made of 50 healthy controls (negatives = 50%) and 50 sick patients (positives = 50%) (Additional file 2c). Imagine that the machine learning prediction generated the following confusion matrix: TP=47, FN=3, TN=5, FP=45 (Additional file 2b).

Once again, the algorithm exhibited its ability to predict the positive data instances (47 sick patients out of 50 were correctly predicted), but it also demonstrated its lack of talent in identifying healthy individuals (only 5 healthy controls of 50 were correctly recognized) (Additional file 2b). Again, the overall performance should be considered mediocre.

Checking only F_1, one would read a good value (0.66 in the [0, 1] interval), and would be overall satisfied about the prediction (Additional file 2a). Once again, this score would hide the truth: the classification algorithm has performed poorly on the negative subset. The Matthews correlation coefficient, instead, by showing a score close to random guessing (+0.07 in the [−1, +1] interval) would be able to inform that the machine learning method has been on the wrong track. Also, it is worth noticing that accuracy would provide with an informative result in this case (0.52 in the [0, 1] interval).

Balanced dataset — Use case B2. As another example, imagine the classifier produced the following confusion matrix: TP = 10, FN = 40, TN = 46, FP = 4 (Additional file 3b).

Similar to what happened for the Use case A2, the method was able to correctly predict many negative cases (46 healthy individuals out of 50), but failed in predicting most of positive data instances (only 10 sick patients were correctly predicted out of 50). Like for the Use case A2, accuracy, F_1 and MCC show average or low result scores (*accuracy* = 0.56, F_1 *score* = 0.31, and *MCC* = +0.17), correctly informing you about the non-optimal performance of the prediction method (Additional file 3a).

Negatively imbalanced dataset — Use case C1. As another example, analyze now this imbalanced dataset made of 90 healthy controls (negatives = 90%) and 10 sick patients (positives = 10%) (Additional file 4c).

Assume the classifier prediction produced this confusion matrix: TP = 9, FN = 1, TN = 1, FP = 89 (Additional file 4b).

In this case, the method revealed its ability to predict positive data instances (9 sick patients out of 10 were correctly predicted), but it also has shown its lack of skill in identifying negative cases (only 1 healthy individual out of 90 was correctly recognized) (Additional file 4c). Again, the overall performance should be judged modest.

Similar to the Use case A2 and *B2*, all three statistical scores generated low results that reflect the mediocre quality of the prediction: F_1 *score* = 0.17 and *accuracy* = 0.10 in the [0, 1] interval, and *MCC* = −0.19 in the [−1, +1] interval (Additional file 4a).

Negatively imbalanced dataset — Use case C2. As a last example, suppose you obtained this alternative confusion matrix, through another prediction: TP = 2, FN = 9, TN = 88, FP = 1 (Additional file 5b).

Similar to the Use case A1 and B1, the method was able to correctly identify multiple negative data instances (88 healthy patients out of 89), but unable to correctly predict most of sick patients (only 2 true positives out of 11 possible elements).

Here, accuracy showed a high value: 0.90 in the [0, 1] interval.

On the contrary, if one decided to take a look at F_1 and at the Matthews correlation coefficient, by noticing low values value (F_1 *score* = 0.29 in the [0, 1] interval and *MCC* = +0.31 in the [−1, +1] interval), she/he would be correctly informed about the low quality of the prediction (Additional file 5a).

As we explained earlier, the key advantage of the Matthews correlation coefficient is that it generates a high quality score *only* if the prediction correctly classified a high percentage of negative data instances *and* a high percentage of positive data instances, with any class balance or imbalance.

Recap. We recap here the results obtained for the six use cases (Table 4). For the Use case A1 (negatively imbalanced dataset), the machine learning classifier was unable to correctly predict negative data instances, and it therefore produced confusion matrices featuring few true negatives (TN). There, accuracy and F_1 generated overoptimistic and inflated results, while the Matthews correlation coefficient was the only statistical rate which identified the aforementioned prediction problem, and therefore to provide a low truthful quality score.

In the Use case A2 (positively imbalanced dataset), instead, the method did not predict correctly enough positive data instances, and therefore showed few true positives. Even if accuracy showed an excessively high result score, the values of F_1 and MCC correctly reflected the low quality of the prediction.

In the Use case B1 (balanced dataset), the machine learning method was unable to correctly predict negative data instances, and therefore produced a confusion matrix featuring few true negatives (TN). In this case, F_1 generated an overoptimistic result, while accuracy and the MCC correctly produced low results that highlight an issue in the prediction.

The classifier did not find enough true positives for the Use case B2 (balanced dataset), too. In this case, all the analyzed rates (accuracy, F_1, and MCC) produced average or low results which correctly represented the prediction issue.

Table 4 Recap of the six use cases results

	Balance		Confusion matrix				Accuracy [0, 1]	F_1 score [0, 1]	MCC [−1, +1]	Figure	Informative Response
	Pos	Neg	TP	FN	TN	FP					
Use case A1 Positively imbalanced dataset	91	9	90	1	0	9	0.90	0.95	**−0.03**	Figure 2	**MCC**
Use case A2 Positively imbalanced dataset	75	25	5	70	19	6	**0.24**	**0.12**	**−0.24**	Suppl. Additional file 1	**Accuracy, F_1 score, MCC**
Use case B1 Balanced dataset	50	50	47	3	5	45	**0.52**	0.66	**+0.07**	Suppl. Additional file 2	**Accuracy, MCC**
Use case B2 Balanced dataset	50	50	10	40	46	4	**0.56**	**0.31**	**+0.17**	Suppl. Additional file 3	**accuracy, F_1 score, MCC**
Use case C1 Negatively imbalanced dataset	10	90	9	1	1	89	**0.10**	**0.17**	**−0.19**	Suppl. Additional file 4	**accuracy, F_1 score, MCC**
Use case C2 Negatively imbalanced dataset	11	89	2	9	88	1	0.90	**0.29**	**+0.31**	Suppl. Additional file 5	**F_1 score, MCC**

For the Use case A1, MCC is the only statistical rate able to truthfully inform the readership about the poor performance of the classifier. For the Use case B1, MCC and accuracy are able to inform about the poor performance of the classifier in the prediction of negative data instances, while for the Use case A2, B2, C1, all the three rates (accuracy, F_1, and MCC) are able to show this information. For the Use case C2, the MCC and F_1 are able to recognize the weak performance of the algorithm in predicting one of the two original dataset classes. pos: number of positives. neg: number of negatives. TP: true positives. FN: false negatives. TN: true negatives. FP: false positives. Informative response: list of confusion matrix rates able to reflect the poor performance of the classifier in the prediction task. We highlighted in bold the informative response of each use case

Also in the Use case C1 (positively imbalanced dataset), the machine learning method was unable to correctly recognize negative data instances, and therefore produced a confusion matrix with a low number of true negative (TN). Here, accuracy again generated an overoptimistic inflated score, while F_1 and the MCC correctly produced low results that indicated a problem in the prediction process.

Finally, in the last Use case C2 (positively imbalanced dataset), the prediction technique failed in predicting negative elements, and therefore its confusion matrix showed a low percentage of true negatives. Here accuracy again generated overoptimistic, misleading, and inflated high results, while F_1 and MCC were able to produce a low score that correctly reflected the prediction issue.

In summary, even if F_1 and accuracy results were able to reflect the prediction issue in some of the six analyzed use cases, the Matthews correlation coefficient was the only score which correctly indicated the prediction problem in all six examples (Table 4).

Particularly, in the Use case A1 (a prediction which generated many true positives and few true negatives on a positively imbalanced dataset), the MCC was the only statistical rate able to truthfully highlight the classification problem, while the other two rates showed misleading results (Fig. 2).

These results show that, while accuracy and F_1 score often generate high scores that do not inform the user about ongoing prediction issues, the MCC is a robust, useful, reliable, truthful statistical measure able to correctly reflect the deficiency of any prediction in any dataset.

Genomics scenario: colon cancer gene expression

In this section, we show a real genomics scenario where the Matthews correlation coefficient result being more informative than accuracy and F_1 score.

Dataset. We trained and applied several machine learning classifiers to gene expression data from the microarray experiments of colon tissue released by Alon et al. [103] and made it publically available within the Partial Least Squares Analyses for Genomics (*plsgenomics*) R package [104, 105]. The dataset contains 2,000 gene probsets for 62 patients, of which 22 are healthy controls and 40 have colon cancer (35.48% negatives and 64.52% positives) [106].

Experiment design. We employed machine learning binary classifiers to predict patients and healthy controls in this dataset: gradient boosting [107], decision tree [108], *k*-nearest neighbors (*k*-NN) [109], support vector machine (SVM) with linear kernel [7], and support vector machine with radial Gaussian kernel [7].

For gradient boosting and decision tree, we trained the classifiers on a training set containing 80% of randomly selected data instances, and test them on the test set containing the remaining 20% data instances. For *k*-NN and SVMs, we split the dataset into training set (60% data instances, randomly selected), validation set (20% data instances, randomly selected), and the test set (remaining 20% data instances). We used the validation set for the hyper-parameter optimization grid search [97]: number *k* of neighbors for *k*-NN, and cost *C* hyper-parameter for the SVMs. We trained each model having a different hyper-parameter on the training set, applied it to the validation set, and then picked the one obtaining the highest MCC as final model to be applied to the test set. For all the classifiers, we repeated the experiment execution ten times and recorded the average results for MCC, F_1 score, accuracy, true positive (TP) rate, and true negative (TN) rate.

We then ranked the results obtained on the test sets or the validation sets first based on the MCC, then based on the F_1 score, and finally based on the accuracy (Table 5).

Results: different metric, different ranking. The three rankings we employed to report the same results (Table 5) show two interesting aspects. First, the top classifier changes when we consider the ranking based on MCC, F_1 score, or accuracy. In the MCC ranking, in fact, the top performing method is gradient boosting (MCC = +0.55), while in the F_1 score ranking and in the accuracy ranking the best classifier resulted being *k*-NN (F_1 score = 0.87 and accuracy = 0.81). The ranks of the other methods change, too: linear SVM is ranked forth in the MCC ranking and in the accuracy ranking, but ranked second in the F_1 score ranking. Decision tree changes its position from one ranking to another, too.

Table 5 Colon cancer prediction rankings

Classifier	MCC	F_1 score	Accuracy	TP rate	TN rate
MCC ranking:					
Gradient boosting	**+0.55**	0.81	0.78	0.85	0.69
Decision tree	**+0.53**	0.82	0.77	0.88	0.58
k-nearest neighbors	**+0.48**	0.87	0.80	0.92	0.52
Linear SVM	**+0.41**	0.82	0.76	0.86	0.53
Radial SVM	**+0.29**	0.75	0.67	0.86	0.40
F_1 score ranking:					
k-nearest neighbors	+0.48	**0.87**	0.80	0.92	0.52
Linear SVM	+0.41	**0.82**	0.76	0.86	0.53
Decision tree	+0.53	**0.82**	0.77	0.88	0.58
Gradient boosting	+0.55	**0.81**	0.78	0.85	0.69
Radial SVM	+0.29	**0.75**	0.67	0.86	0.40
Accuracy ranking:					
k-nearest neighbors	+0.48	0.87	**0.80**	0.92	0.52
Gradient boosting	+0.55	0.81	**0.78**	0.85	0.69
Decision tree	+0.53	0.82	**0.77**	0.88	0.58
Linear SVM	+0.41	0.82	**0.76**	0.86	0.53
Radial SVM	+0.29	0.75	**0.67**	0.86	0.40

Prediction results on colon cancer gene expression dataset, based on MCC, F_1 score, and accuracy. linear SVM: support vector machines with linear kernel. MCC: worst value −1 and best value +1. F_1 score, accuracy, TP rate, and TN rate: worst value 0 and best value 1. To avoid additional complexity and keep this table simple to read, we prefered to exclude the standard deviation of each result metric. We highlighted in bold the ranking of each rate

As mentioned earlier, for binary classifications like this, we prefer to focus on the ranking obtained by the MCC, because this rate generates a high score only if the classifier was able to correctly predict the majority of the positive data instances and the majority of the negative data instances. In our example, in fact, the top MCC ranking classifier gradient boosting did quite well both on the recall (TP rate = 0.85) and on the specificity (TN rate = 0.69). *k*-NN, that is the top performing method both in the F_1 score ranking and in the accuracy ranking, instead, obtained an excellent score for recall (TP rate = 0.92) but just sufficient on the specificity (TN rate = 0.52).

The F_1 score ranking and the accuracy ranking, in conclusion, are hiding this important flaw of the top classifier: *k*-NN was unable top correctly predict a high percentage of patients. The MCC ranking, instead, takes into account this information.

Results: F_1 score and accuracy can mislead, but MCC does not. The second interesting aspect of the results we obtained relates to the radial SVM (Table 5). If a researcher decided to evaluate the performance of this method by observing only the F_1 score and the accuracy, she/he would notice good results (F_1 score = 0.75 and accuracy = 0.67) and might be satisfied about them. These results, in fact, mean 3/4 correct F_1 score and 2/3 correct accuracy.

However, these values of F_1 score and accuracy would mislead the researcher once again: with a closer look to the results, one can notice that the radial SVM has performed poorly on the true negatives (TN rate = 0.40), by correctly predicting less than half patients. Similar to the synthetic Use case A1 previously described (Fig. 2 and Table 4), the Matthews correlation coefficient is the only aggregate rate highlighting the weak performance of the classifier here. With its low value (MCC = +0.29), the MCC informs the readers about the poor general outcome of the radial SVM, while the accuracy and F_1 score show misleading values.

Conclusions

Scientists use confusion matrices to evaluate binary classification problems; therefore, the availability of a unified statistical rate that is able to correctly represent the quality of a binary prediction is essential. Accuracy and F_1

score, although popular, can generate misleading results on imbalanced datasets, because they fail to consider the ratio between positive and negative elements. In this manuscript, we explained the reasons why Matthews correlation coefficient (MCC) can solve this issue, through its mathematical properties that incorporate the dataset imbalance and its invariantness for class swapping. The criterion of MCC is intuitive and straightforward: to get a high quality score, the classifier has to make correct predictions both on the majority of the negative cases, and on the majority of the positive cases, independently of their ratios in the overall dataset. F_1 and accuracy, instead, generate reliable results only when applied to balanced datasets, and produce misleading results when applied to imbalanced cases. For these reasons, we suggest all the researchers working with confusion matrices to evaluate their binary classification predictions through the MCC, instead of using F_1 score or accuracy.

Regarding the limitations of this comparative article, we recognize that additional comparisons with other rates (such as Cohen's Kappa [70], Cramér's V [37], and K measure [81]) would have provided further information about the role of MCC in binary classification evaluation. We prefered to focus on accuracy and F_1 score, instead, because accuracy and F_1 score are more commonly used in machine learning studies related to biomedical applications.

In the future, we plan to investigate further the relationship between MCC and Cohen's Kappa, Cramér's V, K measure, balanced accuracy, F macro average, and F micro average.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6413-7.

Additional file 1: Use case A2 — Positively imbalanced dataset. **(a)** Barplot representing *accuracy*, *F_1 score*, and normalized *Matthews correlation coefficient* ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case A2 positively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 2: Use case B1 — Balanced dataset. **(a)** Barplot representing *accuracy*, *F_1 score*, and normalized *Matthews correlation coefficient* ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case B1 balanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 3: Use case B2 — Balanced dataset. **(a)** Barplot representing *accuracy*, *F_1 score*, and normalized *Matthews correlation coefficient* ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case B2 balanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 4: Use case C1 — Negatively imbalanced dataset. **(a)** Barplot representing *accuracy*, *F_1 score*, and normalized *Matthews correlation coefficient* ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case C1 negatively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 5: Use case C2 — Negatively imbalanced dataset. **(a)** Barplot representing *accuracy*, *F_1 score*, and normalized *Matthews correlation coefficient* ($normMCC = (MCC + 1) / 2$), all in the [0, 1] interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case C2 negatively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Abbreviations

k-NN: *k*-nearest neighbors; AUC: Area under the curve; FDA: Food and drug administration; FN: False negatives; FP: false positives. MAQC/SEQC: MicroArray II / sequencing quality control; MCC: Matthews correlation coefficient; PCC: Pearson correlation coefficient; PLS: Partial least squares; PR: Precision-recall; ROC: Receiver operating characteristic; SVM: Support vector machine; TN: True negatives; TP: True positives

Acknowledgments

The authors thank Julia Lin (University of Toronto) and Samantha Lea Wilson (Princess Margaret Cancer Centre) for their help in the English proof-reading of this manuscript, and Bo Wang (Peter Munk Cardiac Centre) for his helpful suggestions.

Authors' contributions

DC conceived the study, designed and wrote the "Use cases" section, designed and wrote the "Genomics scenario: colon cancer gene expression" section, reviewed and approved the complete manuscript. GJ designed and wrote the "Background" and the "Notation and mathematical foundations" sections, reviewed and approved the complete manuscript. Both the authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data and the R software code used in this study for the tests and the plots are publically available at the following web URL: https://github.com/davidechicco/MCC.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no competing interests.

Author details

[1]Krembil Research Institute, Toronto, Ontario, Canada. [2]Peter Munk Cardiac Centre, Toronto, Ontario, Canada. [3]Fondazione Bruno Kessler, Trento, Italy.

Received: 24 May 2019 Accepted: 18 December 2019
Published online: 02 January 2020

References

1. Chicco D, Rovelli C. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PLoS ONE. 2019;14(1):0208737.

2. Fernandes K, Chicco D, Cardoso JS, Fernandes J. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. PeerJ Comput Sci. 2018;4:154.
3. Maggio V, Chierici M, Jurman G, Furlanello C. Distillation of the clinical algorithm improves prognosis by multi-task deep learning in high-risk neuroblastoma. PLoS ONE. 2018;13(12):0208924.
4. Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, Furlanello C. Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics. 2018;19(2):49.
5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
6. Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4(2):1883.
7. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl. 1998;13(4):18–28.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of KDD 2016 – the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
10. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. Front Biosci. 2008;13:691.
11. Nicodemus KK, Malley JD. Predictor correlation impacts machine learning algorithms: implications for genomic studies. Bioinformatics. 2009;25(15):1884–90.
12. Karimzadeh M, Hoffman MM. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. bioRxiv. 2018;168419:.
13. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48(5):488.
14. Ng KLS, Mishra SK. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics. 2007;23(11):1321–30.
15. Demšar J. Statistical comparisons of classifiers over multiple data sets,. J Mach Learn Res. 2006;7:1–30.
16. García S, Herrera F. An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. J Mach Learn Res. 2008;9:2677–94.
17. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Informa Process Manag. 2009;45:427–37.
18. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recogn Lett. 2009;30:27–38.
19. Garcia V, Mollineda RA, Sanchez JS. Theoretical analysis of a performance measure for imbalanced data. In: Proceedings of ICPR 2010 – the IAPR 20th International Conference on Pattern Recognition. IEEE; 2010. p. 617–20. https://doi.org/10.1109/icpr.2010.156.
20. Choi S-S, Cha S-H. A survey of binary similarity and distance measures. J Syst Cybernet Informa. 2010;8(1):43–8.
21. Japkowicz N, Shah M. Evaluating Learning Algorithms: A Classification Perspective. Cambridge: Cambridge University Press; 2011.
22. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol. 2011;2(1):37–63.
23. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13(4):2.
24. Shin SJ, Kim H, Han S-T. Comparison of the performance evaluations in classification. Int J Adv Res Comput Commun Eng. 2016;5(8):441–4.
25. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv (CSUR). 2016;49(2):31.
26. Ballabio D, Grisoni F, Todeschini R. Multivariate comparison of classification performance measures. Chemom Intell Lab Syst. 2018;174:33–44.
27. Tharwat A. Classification assessment methods. Appl Comput Informa. 20181–13. https://doi.org/10.1016/j.aci.2018.08.003.
28. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recogn. 2019;91:216–31.
29. Anagnostopoulos C, Hand DJ, Adams NM. Measuring Classification Performance: the *hmeasure* Package. Technical report, CRAN. 20191–17.
30. Parker C. An analysis of performance measures for binary classifiers. In: Proceedings of IEEE ICDM 2011 – the 11th IEEE International Conference on Data Mining. IEEE; 2011. p. 517–26. https://doi.org/10.1109/icdm.2011.21.
31. Wang L, Chu F, Xie W. Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans Comput Biol Bioinforma. 2007;4(1):40–53.
32. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Proceedings of Advances in Artificial Intelligence (AI 2006), Lecture Notes in Computer Science, vol. 4304. Heidelberg: Springer; 2006. p. 1015–21.
33. Gu Q, Zhu L, Cai Z. Evaluation measures of the classification performance of imbalanced data sets. In: Proceedings of ISICA 2009 – the 4th International Symposium on Computational Intelligence and Intelligent Systems, Communications in Computer and Information Science, vol. 51. Heidelberg: Springer; 2009. p. 461–71.
34. Bekkar M, Djemaa HK, Alitouche TA. Evaluation measures for models assessment over imbalanced data sets. J Informa Eng Appl. 2013;3(10):27–38.
35. Akosa JS. Predictive accuracy: a misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum 2017 Conference. Cary, North Carolina: SAS Institute Inc.; 2017. p. 942–2017.
36. Guilford JP. Psychometric Methods. New York City: McGraw-Hill; 1954.
37. Cramér H. Mathematical Methods of Statistics. Princeton: Princeton University Press; 1946.
38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA) Protein Struct. 1975;405(2):442–51.
39. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16(5):412–24.
40. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem. 2004;28(5–6):367–74.
41. The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: a comprehensive study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28(8):827–38.
42. The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nat Biotechnol. 2014;32:903–14.
43. Liu Y, Cheng J, Yan C, Wu X, Chen F. Research on the Matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. Int J Hybrid Informa Technol. 2015;8(1):163–72.
44. Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. Oncotarget. 2017;8(57):97025.
45. Brown JB. Classifiers and their metrics quantified. Mol Inform. 2018;37:1700127.
46. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS ONE. 2017;12(6):0177678.
47. Buckland M, Gey F. The relationship between recall and precision. J Am Soc Inform Sci. 1994;45(1):12–9.
48. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):0118432.
49. Dice LR. Measures of the amount of ecologic association between species ecology. Ecology. 1945;26(3):297–302.
50. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. K Dan Vidensk Sels. 1948;5(4):1–34.
51. van Rijsbergen CJ. Foundations of evaluation. J Doc. 1974;30:365–73.
52. van Rijsbergen CJ, Joost C. Information Retrieval. New York City: Butterworths; 1979.
53. Chinchor N. MUC-4 evaluation metrics. In: Proceedings of MUC-4 – the 4th Conference on Message Understanding. McLean: Association for Computational Linguistics; 1992. p. 22–9.
54. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging. 1994;13(4):716–24.

55. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation. In: Information Retrieval Experiment, Chap. 5. Amsterdam: Butterworths; 1981.
56. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, revisited. Informa Process Manag. 1992;28:467–90.
57. Lewis DD. Evaluating text categorization. In: Proceedings of HLT 1991 – Workshop on Speech and Natural Language. p. 312–8. https://doi.org/10.3115/112405.112471.
58. Lewis DD, Yang Y, Rose TG, Li F. RCV1: a new benchmark collection for text categorization research. J Mach Learn Res. 2004;5:361–97.
59. Tsoumakas G, Katakis I, Vlahavas IP. Random k-labelsets for multilabel classification. IEEE Trans Knowl Data Eng. 2011;23(7):1079–89.
60. Pillai I, Fumera G, Roli F. Designing multi-label classifiers that maximize F measures: state of the art. Pattern Recogn. 2017;61:394–404.
61. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: Proceedings of ECML PKDD 2014 – the 2014 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 8725. Heidelberg: Springer; 2014. p. 225–39.
62. Sasaki Y. The truth of the F-measure. Teach Tutor Mater. 2007;1(5):1–5.
63. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8.
64. Powers DMW. What the F-measure doesn't measure...: features, flaws, fallacies and fixes. arXiv:1503.06410. 2015.
65. Van Asch V. Macro-and micro-averaged evaluation measures. Technical report. 20131–27.
66. Flach PA, Kull M. Precision-Recall-Gain curves: PR analysis done right. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015). Cambridge: MIT Press; 2015. p. 838–46.
67. Yedidia A. Against the F-score. 2016. Blogpost: https://adamyedidia.files.wordpress.com/2014/11/f_score.pdf. Accessed 10 Dec 2019.
68. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. Stat Comput. 2018;28:539–47.
69. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. PLoS Comput Biol. 2018;14(12):1006625.
70. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
71. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
72. McHugh ML. Interrater reliability: the Kappa statistic. Biochem Med. 2012;22(3):276–82.
73. Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. Pharm Stat. 2015;14:74–8.
74. Powers DMW. The problem with Kappa. In: Proceedings of EACL 2012 – the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: ACL; 2012. p. 345–55.
75. Delgado R, Tibau X-A. Why Cohen's *Kappa* should be avoided as performance measure in classification. PloS ONE. 2019;14(9):0222916.
76. Ben-David A. Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Syst Appl. 2008;34:825–32.
77. Barandela R, Sánchez JS, Garca V, Rangel E. Strategies for learning in class imbalance problems. Pattern Recogn. 2003;36(3):849–51.
78. Wei J-M, Yuan X-J, Hu Q-H, Wang S-Q. A novel measure for evaluating classifiers. Expert Syst Appl. 2010;37:3799–809.
79. Delgado R, Núñez González JD. Enhancing confusion entropy (CEN) for binary and multiclass classification. PLoS ONE. 2019;14(1):0210264.
80. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE. 2012;7(8):41882.
81. Sebastiani F. An axiomatically derived measure for the evaluation of classification algorithms. In: Proceedings of ICTIR 2015 – the ACM SIGIR 2015 International Conference on the Theory of Information Retrieval. New York City: ACM; 2015. p. 11–20.
82. Espíndola R, Ebecken N. On extending F-measure and G-mean metrics to multi-class problems. WIT Trans Inf Commun Technol. 2005;35:25–34.
83. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceeedings of IAPR 2010 – the 20th IAPR International Conference on Pattern Recognition. IEEE; 2010. p. 3121–4. https://doi.org/10.1109/icpr.2010.764.
84. Dubey A, Tarar S. Evaluation of approximate rank-order clustering using Matthews correlation coefficient. Int J Eng Adv Technol. 2018;8(2):106–13.
85. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
86. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30:1145–59.
87. Flach PA. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: Proceedings of ICML 2003 – the 20th International Conference on Machine Learning. Palo Alto: AAAI Press; 2003. p. 194–201.
88. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17(3):299–310.
89. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
90. Hand DJ. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. Stat Med. 2010;29:1502–10.
91. Suresh Babu N. Various performance measures in binary classification – An overview of ROC study. Int J Innov Sci Eng Technol. 2015;2(9):596–605.
92. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr. 2008;17(2):145–51.
93. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. Bioinformatics. 2010;26(6):822–30.
94. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77(9):103–23.
95. Ferri C, Hernández-Orallo J, Flach PA. A coherent interpretation of AUC as a measure of aggregated classification performance. In: Proceedings of ICML 2011 – the 28th International Conference on Machine Learning. Norristown: Omnipress; 2011. p. 657–64.
96. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. PLoS ONE. 2014;9(3):92209.
97. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min. 2017;10(35):1–17.
98. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68(8):855–9.
99. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 2010;11:523.
100. Sedgwick P. Pearson's correlation coefficient. Br Med J (BMJ). 2012;345:4483.
101. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaest Geographicae. 2011;30(2):87–93.
102. Chicco D, Ciceri E, Masseroli M. Extended Spearman and Kendall coefficients for gene annotation list correlation. In: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer; 2014. p. 19–32. https://doi.org/10.1007/978-3-319-24462-4_2.
103. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci (PNAS). 1999;96(12):6745–50.
104. Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinforma. 2006;8(1):32–44.
105. Boulesteix A-L, Durif G, Lambert-Lacroix S, Peyre J, Strimmer K. Package 'plsgenomics'. 2018. https://cran.r-project.org/web/packages/plsgenomics/index.html. Accessed 10 Dec 2019.
106. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Data pertaining to the article 'Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays'. 2000. http://genomics-pubs.princeton.edu/oncology/affydata/index.html. Accessed 10 Dec 2019.
107. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.
108. Timofeev R. Classification and regression trees (CART) theory and applications. Berlin: Humboldt University; 2004.
109. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? In: International Conference on Database Theory. Springer; 1999. p. 217–35. https://doi.org/10.1007/3-540-49257-7_15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.13 Image analysis and models of cells

Recommended reading for Chapter 13 (listed in Section 13.10).

These papers are recommended for reading in full; most come from the reading lists used in earlier years. Open-access papers are included; the others are available with your SU login through the library link.

- Chamier et al. (2019) — what a biologist should know about deep learning for microscopy, including its pitfalls.
  DOI · open access (CC BY 4.0)

  Artificial intelligence for microscopy: what you should know. von Chamier et al. (2019). *Biochemical Society Transactions*. doi:10.1042/bst20180391

  > Artificial Intelligence based on Deep Learning (DL) is opening new horizons in biomedical research and promises to revolutionize the microscopy field. It is now transitioning from the hands of experts in computer sciences to biomedical researchers. Here, we introduce recent developments in DL applied to microscopy, in a manner accessible to non-experts. We give an overview of its concepts, capabilities and limitations, presenting applications in image segmentation, classification and restoration. We discuss how DL shows an outstanding potential to push the limits of microscopy, enhancing resolution, signal and information content in acquired data. Its pitfalls are discussed, along with the future directions expected in this field.

- Stringer et al. (2020) — Cellpose; a generalist deep-learning method for cell segmentation.
  DOI · via SU library (SU login)

  Cellpose: a generalist algorithm for cellular segmentation. Stringer et al. (2020). *Nature Methods*. doi:10.1038/s41592-020-01018-x

  > Many biological applications require the segmentation of cell bodies, membranes and nuclei from microscopy images. Deep learning has enabled great progress on this problem, but current methods are specialized for images that have large training datasets. Here we introduce a generalist, deep learning-based segmentation method called Cellpose, which can precisely segment cells from a wide range of image types and does not require model retraining or parameter adjustments. Cellpose was trained on a new dataset of highly varied images of cells, containing over 70,000 segmented objects. We also demonstrate a three-dimensional (3D) extension of Cellpose that reuses the two-dimensional (2D) model and does not require 3D-labeled data. To support community contributions to the training data, we developed software for manual labeling and for curation of the automated results. Periodically retraining the model on the community-contributed data will ensure that Cellpose improves constantly.

- Ouyang et al. (2019) — image classification of protein localisation as a community challenge, with lessons for evaluation.
  full text below · DOI · open access (CC BY 4.0)

  Analysis of the Human Protein Atlas Image Classification competition. Ouyang et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0658-6

  > Pinpointing subcellular protein localizations from microscopy images is easy to the trained eye, but challenging to automate. Based on the Human Protein Atlas image collection, we held a competition to identify deep learning solutions to solve this task. Challenges included training on highly imbalanced classes and predicting multiple labels per image. Over 3 months, 2,172 teams participated. Despite convergence on popular networks and training techniques, there was considerable variety among the solutions. Participants applied strategies for modifying neural networks and loss functions, augmenting data and using pretrained networks. The winning models far outperformed our previous effort at multi-label classification of protein localization patterns by ~20%. These models can be used as classifiers to annotate new images, feature extractors to measure pattern similarity or pretrained networks for a wide range of biological applications.

- Chandrasekaran et al. (2024) — Cell Painting and image-based profiling of matched chemical and genetic perturbations (JUMP).
full text below · DOI · open access (CC BY 4.0)

  Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Chandrasekaran et al. (2024). *Nature Methods*. doi:10.1038/s41592-024-02241-6

  > The identification of genetic and chemical perturbations with similar impacts on cell morphology can elucidate compounds' mechanisms of action or novel regulators of genetic pathways. Research on methods for identifying such similarities has lagged due to a lack of carefully designed and well-annotated image sets of cells treated with chemical and genetic perturbations. Here we create such a Resource dataset, CPJUMP1, in which each perturbed gene's product is a known target of at least two chemical compounds in the dataset. We systematically explore the directionality of correlations among perturbations that target the same protein encoded by a given gene, and we find that identifying matches between chemical and genetic perturbations is a challenging task. Our dataset and baseline analyses provide a benchmark for evaluating methods that measure perturbation similarities and impact, and more generally, learn effective representations of cellular state from microscopy images. Such advancements would accelerate the applications of image-based profiling of cellular states, such as uncovering drug mode of action or probing functional genomics.

- Bunne et al. (2024) — the AI virtual cell; what it is meant to be and how it could be built.
full text below · DOI · open access (CC BY 4.0)

  How to build the virtual cell with artificial intelligence: Priorities and opportunities. Bunne et al. (2024). *Cell*. doi:10.1016/j.cell.2024.11.015

  > Cells are essential to understanding health and disease, yet traditional models fall short of modeling and simulating their function and behavior. Advances in AI and omics offer groundbreaking opportunities to create an AI virtual cell (AIVC), a multi-scale, multi-modal large-neural-network-based model that can represent and simulate the behavior of molecules, cells, and tissues across diverse states. This Perspective provides a vision on their design and how collaborative efforts to build AIVCs will transform biological research by allowing high-fidelity simulations, accelerating discoveries, and guiding experimental studies, offering new opportunities for understanding cellular functions and fostering interdisciplinary collaborations in open science.

- Ahlmann-Eltze et al. (2025) — a critical test; deep-learning models of perturbation effects against simple baselines.
full text below · DOI · open access (CC BY 4.0)

  Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Ahlmann-Eltze et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02772-6

  > Recent research in deep-learning-based foundation models promises to learn representations of single-cell data that enable prediction of the effects of genetic perturbations. Here we compared five foundation models and two other deep learning models against deliberately simple baselines for predicting transcriptome changes after single or double perturbations. None outperformed the baselines, which highlights the importance of critical benchmarking in directing and evaluating method development.

- Thornburg et al. (2026) — a mechanistic, spatial model of the complete cell cycle of a minimal cell.
DOI · open access (CC BY 4.0)

  Bringing the genetically minimal cell to life on a computer in 4D. Thornburg et al. (2026). *Cell*. doi:10.1016/j.cell.2026.02.009

  > We present a whole-cell spatial and kinetic model for the ~100 min cell cycle of the genetically minimal bacterium JCVI-syn3A. We simulate the complete cell cycle in 4D (space and time), including all genetic information processes, metabolic networks, growth, and cell division. By integrating hybrid computational methods, we model the dynamics of morphological transformations. Growth is driven by insertion of lipids and membrane proteins and constrained by fluorescence imaging data. Chromosome replication and segregation are controlled by the essential structural maintenance of chromosome proteins, analogous to condensin (SMC) and topoisomerase proteins in Brownian dynamics simulations,

with replication rates responding to deoxyribonucleotide triphosphate (dNTP) pools from metabolism. The model captures the origin-to-terminus ratio measured in our DNA sequencing and recovers other experimental measurements, such as doubling time, mRNA half-lives, protein distributions, and ribosome counts. Because of stochasticity, each replicate cell is unique. We predict not only the average behavior of partitioning to daughter cells but also the heterogeneity among them.

The open-access papers are reproduced below under their original licences.

ANALYSIS | FOCUS
https://doi.org/10.1038/s41592-019-0658-6
nature methods
OPEN
There are amendments to this paper

Analysis of the Human Protein Atlas Image Classification competition

Wei Ouyang[1], Casper F. Winsnes[1], Martin Hjelmare[1], Anthony J. Cesnik[2,3], Lovisa Åkesson[1], Hao Xu[1], Devin P. Sullivan[1], Shubin Dai[4], Jun Lan[5], Park Jinmo[6], Shaikat M. Galib[7], Christof Henkel[8], Kevin Hwang[9], Dmytro Poplavskiy[10], Bojan Tunguz[11], Russel D. Wolfinger[12], Yinzheng Gu[13], Chuanpeng Li[13], Jinbin Xie[13], Dmitry Buslov[14], Sergei Fironov[15], Alexander Kiselev[16], Dmytro Panchenko[17], Xuan Cao[18], Runmin Wei[19], Yuanhao Wu[20], Xun Zhu[21], Kuan-Lun Tseng[22], Zhifeng Gao[23], Cheng Ju[24], Xiaohan Yi[25], Hongdong Zheng[26], Constantin Kappel[27] and Emma Lundberg[1,2,3]*

Pinpointing subcellular protein localizations from microscopy images is easy to the trained eye, but challenging to automate. Based on the Human Protein Atlas image collection, we held a competition to identify deep learning solutions to solve this task. Challenges included training on highly imbalanced classes and predicting multiple labels per image. Over 3 months, 2,172 teams participated. Despite convergence on popular networks and training techniques, there was considerable variety among the solutions. Participants applied strategies for modifying neural networks and loss functions, augmenting data and using pretrained networks. The winning models far outperformed our previous effort at multi-label classification of protein localization patterns by ~20%. These models can be used as classifiers to annotate new images, feature extractors to measure pattern similarity or pretrained networks for a wide range of biological applications.

Advancement in high-throughput microscopy has propelled the generation of massive amounts of biological imaging data[1]. These data can offer valuable insights into cellular processes and biological systems, but only if we conquer the challenges of processing these voluminous datasets. The Human Protein Atlas (HPA) is a project that faces these challenges and opportunities. Through a systematic antibody-based approach, millions of fluorescence microscopy images have been generated to map the expression of the human proteome[2]. Compartmentalization is an important mechanism to allow multiple biological reactions to occur in parallel. In cells, these compartments are called organelles, and knowledge of which compartments our proteins reside in would greatly increase our understanding of human biology. The HPA Cell Atlas aims to map the subcellular distribution of the human proteome with confocal microscopy[3]. The current version (HPAv19) of the database comprises image data for 12,390 proteins (www.proteinatlas.org).

We have previously demonstrated an astounding degree of cellular complexity; as many as half of all human proteins are localized to multiple cellular compartments[3], and many proteins show single-cell variability[4]. Unbiased analysis of subcellular protein localizations from our images has greatly enriched our vocabulary for describing cellular systems. This analysis was first performed manually[3], and we have since integrated the labor-intensive annotation tasks into a mainstream video game[5], which produced tens of millions of human annotations. These annotations were particularly successful at the challenging task of identifying mixed patterns of protein localizations, a task called multi-label classification[6]. Previously, we developed a machine learning model, called Loc-CAT, capable of classifying mixed patterns in images of cell types of different morphology[5]. However, the performance measured with macro F1 score (defined in Methods) was yet substantially lower (0.47) than that of human experts (0.71).

The emerging field of deep learning[7] has powered many successful real-life applications, including image recognition[8], gaming[9] and autonomous cars[10]. Deep neural networks, particularly convolutional neural networks (CNNs)[8], have been widely applied to perform computer vision tasks such as image classification[11,12] and segmentation[13]. Compared to Loc-CAT[5], which uses hand-crafted features as inputs, CNNs typically take raw images as inputs and learn hierarchical feature representations in an end-to-end fashion. This allows the model to better abstract cellular localization patterns and scale efficiently with data size[14]. CNNs are increasingly used for biological image analysis[15–17] including multi-label classification for yeast protein localization[18]. Over recent years, a collection of successful neural network architectures, such as Resnet[19], Inception[12] and Densenet[20], and different training techniques, such as Dropout[21], Batch Normalization[22], Focal Loss[23], Cyclical learning rates[24] and AutoAugment[25], have been developed. Software libraries, such as PyTorch[26] and Tensorflow[27], can be easily implemented and applied to a wide range of applications. Automated machine learning[28](AutoML) techniques such as hyperparameter optimization[29], meta-learning[30] and neural architecture search[31] make the model development easier and accessible even for nonexperts.

Finding the best solution for classifying protein localizations within HPA Cell Atlas Images involves performing searches of

[1]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH–Royal Institute of Technology, Stockholm, Sweden. [2]Department of Genetics, Stanford University, Stanford, CA, USA. [3]Chan Zuckerberg Biohub, San Francisco, CA, USA. [4]Changsha, China. [5]Winning Health Technology Group Co., Ltd., Shanghai, China. [6]Seoul, Republic of South Korea. [7]Missouri University of Science and Technology, Rolla, MO, USA. [8]Khumbu.ai, Munich, Germany. [9]Qualcomm, Inc., Cupertino, CA, USA. [10]Brisbane, Queensland, Australia. [11]H2O.ai, Greencastle, IN, USA. [12]SAS Institute, Inc., Cary, NC, USA. [13]Jilian Technology Group (Video++), Shanghai, China. [14]SAP, Moscow, Russian Federation. [15]BDO Unicon, Saint Petersburg, Russian Federation. [16]Ivanovo, Russian Federation. [17]Kharkiv National University of Radioelectronics, Kharkiv, Ukraine. [18]Santa Clara, CA, USA. [19]UT MD Anderson Cancer Center, Houston, TX, USA. [20]Shanghai, China. [21]University of Hawaii Cancer Center, Honolulu, HI, USA. [22]Taipei, Republic of China. [23]Microsoft Research, Beijing, China. [24]University of California Berkeley, Berkeley, CA, USA. [25]Beijing, China. [26]Peking University, Beijing, China. [27]Leica Microsystems, Mannheim, Germany. *e-mail: emma.lundberg@scilifelab.se

Table 1 | Models and their performance for top ranking and selected teams

Rank	Team Name	Member(s)	Score	Award (US$)
1	Team 1: bestfitting	D. Shubin	0.593	14,000
2	Team 2: WAIR	J. Lan	0.571	10,000
3	Team 3: pudae	P. Jinmo	0.570	8,000
4	Team 4: Wienerschnitzelgemeinschaft	S. Mahmood Galib, C. Henkel, K. Hwang, D. Poplavskiy, B. Tunguz, R. Wolfinger	0.567	5,000
5	Team 5: vpp	Y. Gu, C. Li, J. Xie	0.566	-
8	Team 8: One More Layer (Of Stacking)	D. Buslov, S. Fironov, A. Kiselev, D. Panchenko	0.563	-
10	Team 10: conv is all u need	X. Cao, R. Wei, Y. Wu, X. Zhu	0.557	-
16	Team 16: NTU_MiRA	K.-L. Tseng	0.553	-
39	Team 39: Random Walk	Z. Gao, C. Ju, X. Yi, H. Zheng	0.540	-

parameters and hyperparameters over an enormous solution space. Crowd-sourced competitions are commonly used for large-scale solution searches. One such success is ILSVRC[32], which provides the ImageNet dataset[33] that is widely recognized as powering the current advancements in deep learning. Competitions for cellular image analysis have also been successfully introduced; for example, for classification[34] and cell-tracking[35]. Popular online platforms such as Kaggle, Innocentive and DREAM allow publishing datasets and hosting competitions typically focused on benchmarking methods for fundamental research problems (for example, DREAM) or crowdsourcing solutions for real-life applications, often motivated by considerable prize money (for example, Kaggle).

Here, we present the design and results from our 'Human Protein Atlas Image Classification' competition, hosted by Kaggle. In contrast to typical image classification tasks that predict one label per image, our dataset requires classification of multiple labels per image (the multi-label problem[6]). There is also a great class imbalance in the dataset, making the classification task harder (the class imbalance problem[36]). During a 3-month period, 2,172 teams provided a total of 55,213 submissions, nearly all based on deep learning. The top-ranking solutions were awarded with a cash prize (Table 1, first place, US$14,000; second place, US$10,000; third place, US$8,000; fourth place: US$5,000). The models far outperformed our previous effort at protein localization classification, and there was considerable variety among the solutions with some convergence on popular networks and training techniques. Different strategies for adapting neural networks and loss functions, augmenting data and using pretrained networks were successfully applied by the winning teams. Here, we present the competition design, statistical analyses of the solutions and visualizations of the winning models to shed light on the considerations for designing multi-label pattern classification algorithms and potential applications of the winning solutions.

Results

Competition design and assessment metrics. The aim of the competition was to develop computational models for the classification of protein subcellular localization patterns in confocal microscope images from the HPA Cell Atlas (Fig. 1). We prepared a dataset of 42,774 nonpublic images and allowed participants to use any external data, including the ~78,000 images that are publicly available on the HPA Cell Atlas (HPAv18). Each image contained multiple cells and had four channels marking the protein of interest and cell outlines (Fig. 1a).

We designed the competition to address the two main challenges of developing computational models for this purpose, namely the class imbalance and multi-label problems (Fig. 1). The first difficulty arises from the highly imbalanced frequencies of the 28 localization classes (Fig. 2a and Supplementary Table 1). The most common label in the training set was 'nucleoplasm' (12,885 images in training and 31,590 images in HPAv18), while the rarest label was 'rods and rings' (11 images in training and 42 images in HPAv18, see Supplementary Table 2). The second difficulty, the multi-label problem, stems from the need to potentially assign multiple labels to each image. Each image has been assigned 1–6 such labels during the standardized annotation pipeline[3] (Fig. 1c, Methods). In total, the dataset contains 577 unique combinations of labels. Beyond these main difficulties, different cellular morphologies of the 27 cell lines in our dataset add complexity (Supplementary Table 3).

Typically, learning algorithms perform well on common classes but perform poorly on rare classes. To encourage an equally distributed classification performance among all the 28 classes, we chose to assess model performance with a macro F1 score (range 0–1). This score gives importance to both precision and sensitivity (recall), and is calculated for each class before averaged over the 28 classes. To earn a high score in this competition, a model thus needs to pay special attention to rare classes. During the competition, teams could see their score and rank on a public leaderboard, which was computed from a subset of test images ('validation_public', see Fig. 2a). At the end of the competition, another subset of test images ('test_private', see Fig. 2a) was used to produce the final ranking on the private leaderboard (Methods). During the competition, ~148 images of rare classes were mistakenly included from the published Cell Atlas Images; this was fixed during the competition by excluding the leaked images from the 'test_private' set (full disclosure in Methods). Since the participants were notified and the final evaluation was performed with the 'test_private' set, the overall impact of the leakage on the results of this paper is minimal.

Participation and performance. After the final ranking, we awarded the top four teams based on their rank on the private leaderboard (Table 1). The winning team generated models with macro F1 scores ranging from ~0.56–0.59, >20% better than Loc-CAT and the citizen science results. We invited nine teams among the top 40 to participate in this study.

When analyzing the final submission for all the teams (2,137 submissions), we found that the top teams were marginally better than most other teams. Both the precision and recall of experts are substantially higher than the top teams. Figure 2b shows the precision-recall for all the teams and the experts' score (on a similar dataset; precision, 0.74 and recall, 0.69) from our previous work[5]. We computed macro F1 scores for groups of teams binned based on their ranking on the private leaderboard (Fig. 2d). The scores for teams with higher ranks varied less, which indicates it is harder to improve further for higher ranking models. The scores for single label images are significantly higher (Methods, $P < 1.08 \times 10^{-5}$, two-sample Kolmogorov–Smirnov test) than for multi-label images

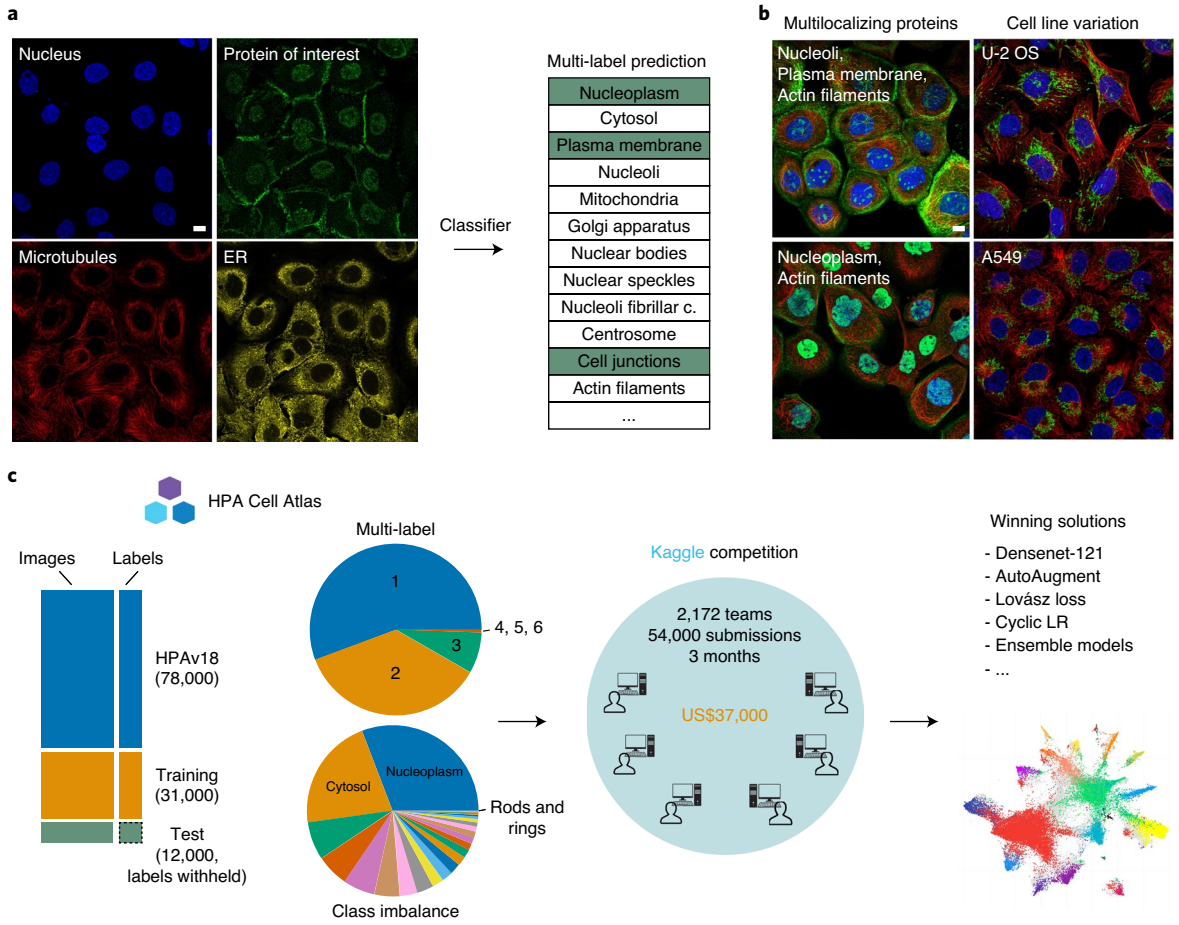


Fig. 1 | Overview of image dataset and challenge design. a, A typical HPA Cell Atlas image and the aim of the competition. Each image consists of four channels: the antibody-stained protein of interest (green) and three reference channels to outline the cell: microtubules (red), nucleus (blue) and endoplasmic reticulum (ER; yellow). The human cell comprises many compartments, here defined by 28 labels. The aim of the competition is to build classifiers to predict the localization pattern (often multiple labels) of the protein of interest. Scale bar, 10 µm. **b**, Sample images showing different protein or cell line expression patterns that make the pattern classification task challenging. Proteins localizing to multiple compartments are exemplified by Septin 7 in A-431 cells (left, top), and PRAME family member 12 in A-431 cells (left, bottom). Stainings of mitochondria (TOMM70, Translocase of outer mitochondrial membrane 70, in U-2 OS and CCR7, C-C motif chemokine receptor 7, in A459) show the morphological differences between cell lines (right, from top to bottom). Scale bar, 10 µm. PM, plasma membrane; Actin fil., actin filaments. **c**, Challenge overview: the HPA is a proteome-wide image collection detailing protein localization. This dataset is challenging to analyze automatically because of prevalent multi-label classifications (1–6 labels per image, upper pie chart) and high imbalance among the 28 different protein localization classes (lower pie chart). To find the best solution for these problems, we held a competition hosted by Kaggle. The challenge dataset consisted of 42,774 images with labels from expert annotations and was divided into a training set and test set before distributed to the Kaggle challenge participants with the labels of the test set withheld. We used a macro F1 score to assess the performance of these models. The competition produced winning solutions and different methods for multi-label image classification. LR, learning rate.

for all groups, emphasizing the difficulty of classifying images with mixed localization patterns.

To understand the impact of class imbalance on the performance of the models, we computed the class-wise score distribution for the top ten teams. Figure 2c and Supplementary Table 4 show that the averaged F1 score, precision and recall for each class varied more when the class had fewer samples in the training dataset; in particular, most models struggled with the two rarest classes (rods and rings, microtubule ends). The difficulty of identifying different localization patterns in the images also played a role in performance for each class. For example, the average F1 score was higher for aggresomes than plasma membrane, despite a lower number of training images (322 and 3,777, respectively), because the aggresome is visually distinct while the plasma membrane is often confused with the cytosol. The performance correlated better with the sample number for cell lines than for classes (Supplementary Fig. 1), likely because the assessment metric (macro F1) encouraged participants to develop models that are equal in performance for the different classes, rather than the different cell lines. A confusion matrix (Supplementary Fig. 2) for single-class samples based on the winning model shows that the endoplasmic reticulum and peroxisomes were confused with cytosol, and that nuclear speckles were mistakenly classified as nucleoplasm. The patterns are consistent across cell lines, despite morphological differences, showing that the model generalizes well (Supplementary Fig. 2).

Strategies used by the top-ranking solutions. To compare the underlying structures of the solutions, we invited the top 200 teams to fill out a survey on the methodology used, which was answered by 56 teams. Notably, all teams but one used deep learning models. For the neural network architectures, 44 of the 56 teams used variations of Resnet, Densenet or Inception as backbone architecture, as they are known to be effective for image recognition tasks[12,19,20]. To address the multi-label problem, most teams (34 of 56) used binary

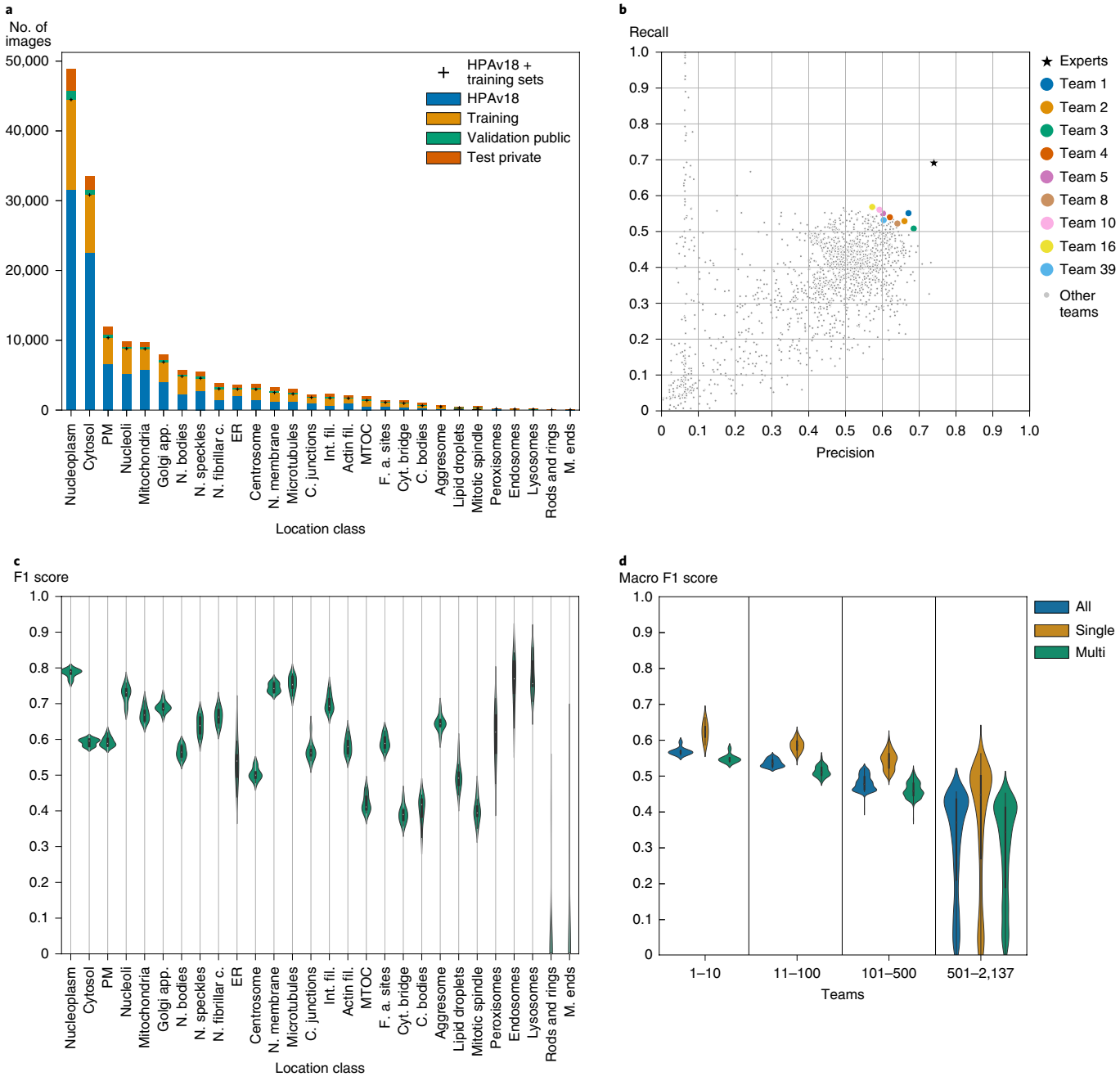


Fig. 2 | Competition results. a, Image numbers of each localization class for HPAv18, training, validation_public and test_private dataset. PM, plasma membrane; Golgi app., Golgi apparatus; N. bodies, nuclear bodies; N. speckles, nuclear speckles; N. fibrillar c., nucleolar fibrillar center; ER, endoplasmic reticulum; N. membrane, nuclear membrane; C. junctions, cell junctions; Int. fil., intermediate filaments; Actin fil., actin filaments; MTOC, microtubule organizing center; F. a. sites, focal adhesion sites; Cyt. bridge, cytokinetic bridge; C. bodies, cytoplasmic bodies; M. ends, mitochondrial ends. **b**, Precision-recall values for the experts, selected teams (including the top four winning teams) and all other teams. **c**, Statistics on the macro F1 scores of different teams and their performance on different classes. Score distributions for the different label classes with the classes sorted according to sample size (high to the left, low to the right). $n = 10$ teams for each violin. The minimum (min), mean, percentile (P) and maximum (max) values can be found in Supplementary Table 9. **d**, Statistics on the macro F1 scores of different teams and their performance, binned into groups based on their ranking on the leaderboard. The top 10, 11–100, 101–500 and the remaining teams, respectively. The scores for single localized, multi-localized and all proteins are shown separately. $n = 10$ teams for violins with teams 1–10, $n = 90$ teams for violins with teams 11–100, $n = 400$ teams for violins with teams 101–500 and $n = 1{,}637$ teams for violins with teams 501–2,137. The minimum, mean, percentile and maximum values can be found in Supplementary Table 9.

cross entropy. Many teams handled class imbalance by applying class weights or by using focal loss[23] to train the models, and employed multi-label stratification[37] to generate validation datasets. Most teams used the public HPAv18 dataset for training; adding these ~78,000 annotated images led to a substantial boost in image classification scores (for example, from 0.510 to 0.552 for Team 1, see Supplementary Tables 5b and 6), mostly due to the increase of rare class images. Augmentation strategies such as random cropping, rotation and flipping were commonly used (Supplementary Notes and Supplementary Table 5).

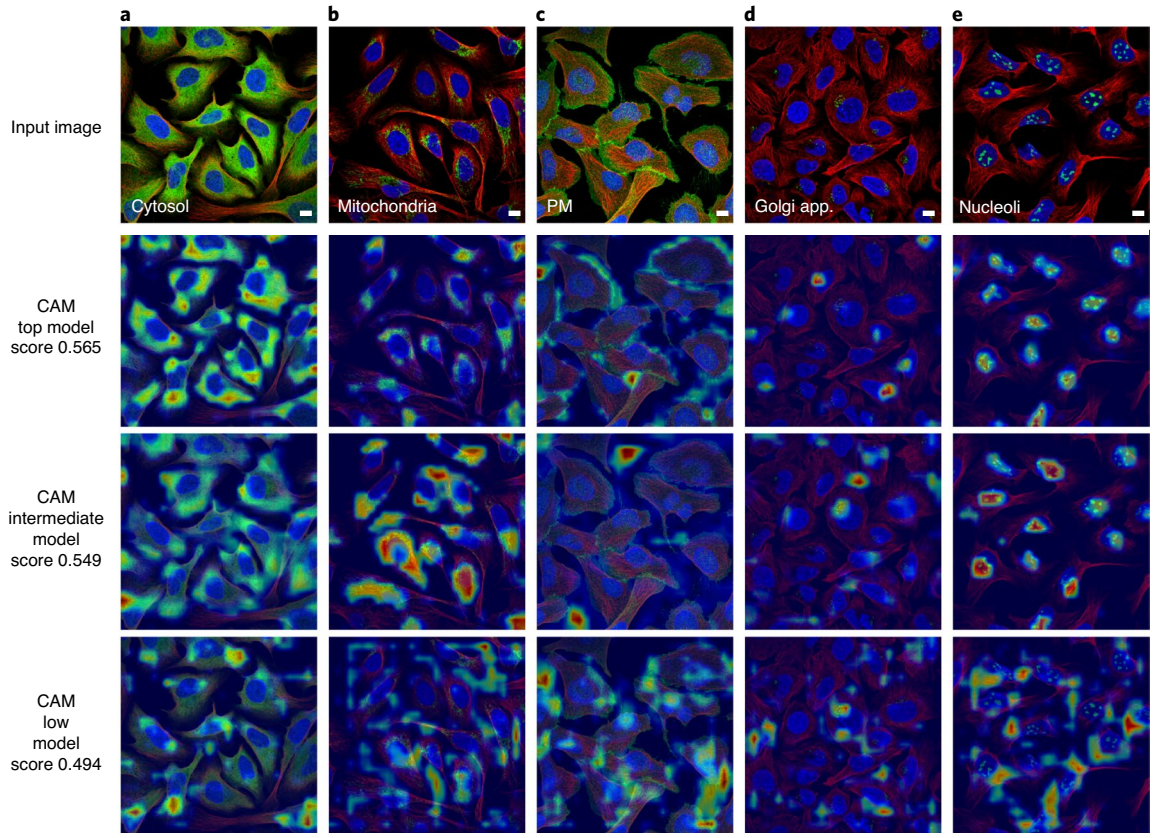


Fig. 3 | Visualization of model spatial attention. CAMs for three different models, the top-scoring model (from Team 1), an intermediate-scoring model (from Team 3) and a low-scoring model (from Team 1). Scale bars, 10 µm. **a**, For the cytosolic protein Methenyltetrahydrofolate synthetase, the CAMs for all three models highlight relevant cellular regions. **b**, The CAMs for the mitochondrial protein Prohibitin 2 show a progressively worse overlap with the mitochondrial staining following the model accuracy score. **c**, The plasma membrane staining of Catenin beta 1 overlaps well with the CAM for the top model, but not for the intermediate and lower scoring models. **d**, The CAMs for Golgi reassembly stacking protein 1, which is localized to the Golgi apparatus, show attention of correct size for all three models, but none of the models focused on all cells in the image. **e**, The nucleolar staining pattern of UTP6 small subunit processome component, is captured well by the CAMs for the top and intermediate models in the nuclear region of the cell.

To understand how these different strategies contribute to performance, we collected detailed information about the solutions and intermediate experiments from nine selected teams (Table 1, Supplementary Table 5 and Supplementary Figs. 3–11). Among these teams, different strategies were applied to different aspects of image analysis. For example, Team 1 used an optimized single neural network and a combined loss function with a Lovász loss[38] term, Team 2 focused on data preprocessing, Team 3 used automatic data augmentation and Team 4 ensembled a large number of models.

We found that different teams often drew the same conclusion on a number of strategies, despite the fact that they mostly worked independently. Team 3 employed an automated augmentation strategy search algorithm (AutoAugment[25]), which resulted in an improvement of the macro F1 score from 0.477 to 0.499. In addition to training time augmentations, test time augmentations were also shown to be effective by several teams. Both Team 1 and 5 found DenseNet[20] to be more effective than Resnet[19] in terms of neural network architecture. In terms of network size, Teams 1 and 4 found medium-sized networks (for example, Densenet 121) to be better than larger ones (for example, Densenet 169). Several teams, including 1, 2, 3 and 16, found that scores can be improved by using a larger image size (for example, 1,024 × 1,024 pixels). Teams 4 and 10 even applied models that work on multi-scale. Techniques such as model ensembling and stacking were used to push the performance beyond single models. By ensembling single models from the top three teams, we obtained an even better model (macro F1 = 0.575, Supplementary Table 7).

To facilitate the reuse of these solutions, we built a model zoo (https://modelzoo.cellprofiling.org) to host the source code and trained models from the selected teams. These can be used as pretrained models[39] to reduce the training time and training data size for constructing new models for biological image analysis.

Assessing the biological relevance of the winning model with class activation maps (CAMs). Understanding whether a prediction for a given image is based on biologically relevant information is difficult due to the poor interpretability of neural networks. However, newly developed visualization techniques, such as class activation mapping[40], allow us to peer into the spatial attention of these models to ensure that the classification is based on biologically meaningful information. A model that fails to focus on the biologically relevant regions of a cell generally indicates lower performance. In Fig. 3, we compared the best, an intermediate, and a low scoring model (Methods) by generating CAMs from 'easy' (cytosol and nucleoli) and 'hard' patterns (mitochondria, plasma membrane, Golgi apparatus). For the easy patterns, the CAMs visualize biologically relevant regions for the top and intermediate model. For the hard patterns, we see more diversity between the models and images.

CAMs can also be used to identify when the score fails to reflect biological information. The final solution by Team 1 is an ensemble of two models (Supplementary Fig. 3): a classification model and a metric learning model[41] that was trained with antibody identifiers from HPAv18 (Supplementary Notes, Supplementary Fig. 12 and Supplementary Table 5). The metric learning part boosted the score

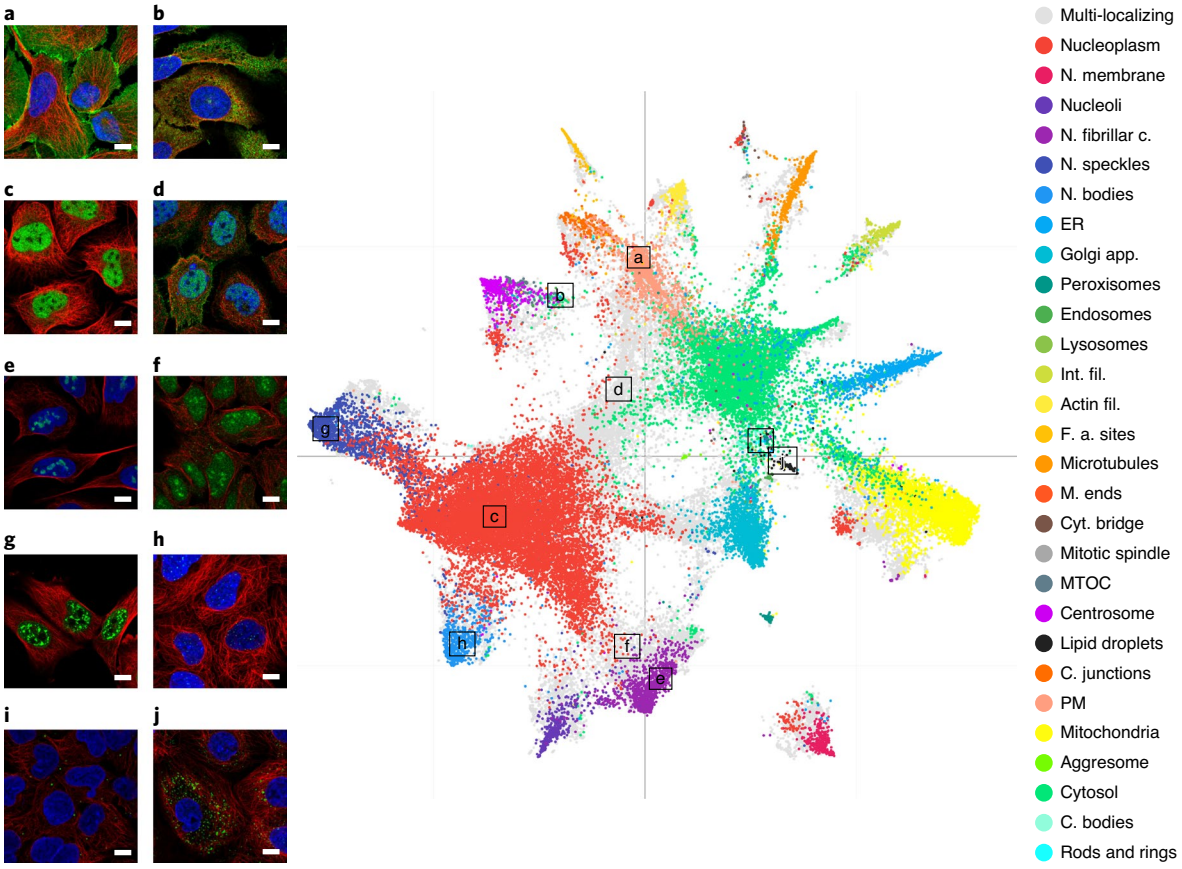


Fig. 4 | Visualization of learned features. UMAP visualization of the features learned by the best scoring model from Team 1 with a few corresponding original images highlighted. Single location images are colored according to location, while gray data points belong to multi-localizing proteins. Abbreviations as in Fig. 2. Scale bars, 10 µm. **a**, Catenin beta 1 is localized to the plasma membrane and also appears in the plasma membrane protein cluster. **b**, Although trained on the manual labels, this type of unbiased analysis provides a tool to identify misclassified patterns or subtle pattern variations. The protein suppressor of cytokine signaling 3 with the label 'cytosol' is found among the centrosome/microtubule organizing center (MTOC) cluster. After visual inspection, we can indeed identify an enrichment of this protein around the MTOC in addition to the cytoplasm in some cells. **c**, RUNX1 translocation partner 1 is localized to the nucleoplasm and appears in the nucleoplasmic protein cluster. **d**, Utrophin is localized to both the plasma membrane and nucleoplasm and appears between these two respective clusters. **e**, EBNA1 binding protein 2 is localized to nucleoli and appears in the nucleoli cluster. **f**, L3MBTL3 histone methyl-lysine binding protein is localized to both the nucleoli and nucleus, and appears between these two respective clusters. **g**,**h**, Heterochromatin protein 1 binding protein 3 is localized to nuclear speckles (**g**) and Centromere protein T (**h**) is localized to centromeres. Despite the pattern similarities of the two categories, they still appear in two distinct clusters. **i**,**j**, Enhancer of mRNA decapping 4 protein is localized to cytoplasmic bodies (**i**), generating a similar staining pattern as Perilipin 3, which is localized to lipid droplets (**j**). Despite the similarities of the two categories, they still appear in two distinct clusters.

by ~4% (macro F1 of 0.565 with the classification part alone to 0.593). However, we found that the metric learning model mainly gained performance by pairing images in the test set to those acquired from the same sample in HPAv18 (Methods). With the help of CAMs, we found that the visual attention pattern varied from protein to protein and often focused on biologically meaningless features (Supplementary Fig. 13). Presumably, the metric learning model works by exploiting 'batch effects' in the dataset through so called 'hidden variables', a pitfall in machine learning[40]. Nevertheless, the classification model without metric learning retained the record of single model performance (macro F1 = 0.565, Supplementary Table 5) with verified performance in a five-fold cross-validation experiment (Supplementary Table 8).

Visualizing image feature representations of subcellular protein distributions. To investigate the ability of a CNN to distinguish subcellular structures, we analyzed features extracted from the penultimate layer of the classification model from Team 1 (without the metric learning model). Figure 4 shows a uniform manifold approximation and projection for dimension reduction (UMAP)[42] projection of these features. First, the features clearly distinguish the different subcellular locations. Nuclear sub-compartments cluster separately, such as the nucleoplasm (for example, RUNX1 translocation partner 1), nucleoli (for example, EBNA1 binding protein 2), nucleoli fibrillar center, nuclear bodies (for example, Centromere protein T) and nuclear speckles (for example, Heterochromatin protein 1 binding protein 3); these examples also illustrate how well fine structures are distinguished, such as the assignment of heterochromatin protein 1 binding protein to nuclear speckles and centromere protein T to nuclear bodies (that is, centromeres). Similarly, the proteins Enhancer of mRNA decapping 4 and Perilipin 3 are accurately predicted to localize to cytoplasmic bodies and lipid droplets. Although these cellular structures are small puncta in the cytosol, the model can reliably distinguish them.

Images that were assigned multiple localization labels (gray points in Fig. 4) are found between clusters of images representing the single locations. Examples include L3MBTL3 histone methyl-lysine binding protein, located in both the nucleus and nucleoli, and

Utrophin, located in both the nucleus and plasma membrane. This indicates that the features learned by the CNN have the potential to be used as quantitative representations of mixed patterns.

A web application was developed and deployed with the ImJoy platform[43], where users can interact with the UMAP and the images with associated metadata (https://tinyurl.com/y6nhf5bo). This provides a new way to explore large online resources, such as the HPA database.

Discussion

The HPA image classification competition provides crowd-sourced solutions for the task of classifying protein subcellular localization patterns in fluorescence microscopy images. The participants were tasked with developing solutions to solve the multi-label classification problem on a dataset with high class imbalance. The participants tested a large number of techniques in a competitive setup, which led to the use of external datasets and methods not previously applied to multi-label classification.

A key design choice for the competition was to use macro F1 as the assessment metric. It successfully encouraged the participants to optimize their solution to handle the label class imbalance in our dataset and yielded roughly similar performances among the classes except the rarest two. During the challenge, many strategies were implemented, such as adapting Lovász loss (designed for segmentation) to multi-label classification, developing the metric learning model and testing a large number of new models and training techniques. As a result, the performance of the winning models are substantially better than our previous Loc-CAT model, but not yet at the level of human experts. More comparative experiments are required to better evaluate the gap between our models with human experts.

Despite the superior performance of deep learning methods, there are limitations including hallucination and generalization problems[44]. It has also been reported that machine learning algorithms can easily pick up unintentional variations (for example, in biomedical image classification[45]). We speculate that the metric learning model used these unintended hidden variables[46](for example, background noise) to match the test images with HPAv18 and boost the performance. Since this type of problem cannot be reflected by the evaluation metric, special attention is required for future competition organizers to prevent this type of exploitation.

We envision that the pretrained models provided in our model zoo will be useful in the context of transfer learning, a popular method that dramatically reduces training time and improves the generalization of learning models[39]. As common practice, models pretrained with ImageNet[32,33] are used even when applied to microscopy images of cells[47], showing the robustness of transfer learning. However, it has also been shown that the generalization of the ImageNet features is questionable, especially for fine-grained classification tasks[48]. Compared to the classes in ImageNet (for example, house and plane), the patterns in cells (for example, mitochondria and centrosome) are much finer grained. Thus, we foresee that applications involving biological images will adopt the HPA classification models trained with a large number of cell microscopy images instead of using ImageNet. Furthermore, our HPA competition dataset can be used as a benchmark dataset similar to ImageNet for developing new algorithms. For example, it may aid in designing models that handle high class imbalance or advancing cell mapping research by developing models for unsupervised analysis of subcellular protein patterns in single cells.

Although the top model could not reach expert level performance, it still opens up avenues for advancing cell biology. The use of high-dimensional features (as in the UMAP) instead of discrete labels constitutes an attractive approach for building systems level representations of cells harboring information about both single and multi-localizing proteins and serves a basis for quantitative integration of spatial information with other 'omics data.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at https://doi.org/10.1038/s41592-019-0658-6.

Received: 30 April 2019; Accepted: 17 October 2019;
Published online: 28 November 2019

References

1. Ouyang, W. & Zimmer, C. The imaging tsunami: computational opportunities and challenges. *Curr. Opin. Syst. Biol.* **4**, 105–113 (2017).
2. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
3. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
4. Mahdessian, D. et al. Spatiotemporal dissection of the cell cycle regulated human proteome. Preprint at *bioRxiv* https://doi.org/10.1101/543231 (2019).
5. Sullivan, D. P. et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).
6. Tsoumakas, G. & Katakis, I. Multi-label classification: an overview. *Int. J. Data Warehous. Min.* **3**, 1–13 (2009).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
8. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *IEEE*, **86**, 2278–2324 (1998).
9. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
10. Bojarski, M. et al. End to end learning for self-driving cars. Preprint at https://arxiv.org/abs/1604.07316 (2016).
11. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/abs/1409.1556 (2014).
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
13. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (eds Navab, N. et al.) 234–241 (Springer, 2015).
14. Hestness, J. et al. Deep learning scaling is predictable, empirically. Preprint at https://arxiv.org/abs/1712.00409 (2017).
15. Moen, E. et al. Deep learning for cellular image analysis. *Nat. Methods* https://doi.org/10.1038/s41592-019-0403-1 (2019).
16. Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W. & Zhang, X. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinforma. Oxf. Engl.* **33**, 2010–2019 (2017).
17. Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S. & Klambauer, G. accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *J. Chem. Inf. Model.* **59**, 1163–1171 (2019).
18. Kraus, O. Z., Ba, J. L. & Frey, B. J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**, i52–i59 (2016).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
20. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
22. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/abs/1502.03167 (2015).
23. Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. in *IEEE International Conference on Computer Vision* 2980–2988 (IEEE, 2017).
24. Smith, L. N. Cyclical learning rates for training neural networks. in *IEEE Winter Conference on Applications of Computer Vision* 464–472 (IEEE, 2017).
25. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V. AutoAugment: learning augmentation policies from data. Preprint at https://arxiv.org/abs/1805.09501 (2018).

26. Paszke, A. et al. Automatic differentiation in PyTorch. in *NIPS 2017 Autodiff Workshop* (2017).
27. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2016).
28. Hutter, F., Kotthoff, L. Vanschoren, J. *Automated Machine Learning-Methods, Systems, Challenges* (Springer International Publishing, 2019).
29. Falkner, S., Klein, A. & Hutter, F. BOHB: robust and efficient hyperparameter optimization at scale. in *35th International Conference on Machine Learning* 1436–1445 (ICML, 2018).
30. Vanschoren, J. Meta-learning: a survey. Preprint at https://arxiv.org/abs/1810.03548 (2018).
31. Elsken, T., Metzen, J. H. & Hutter, F. Neural architecture search: a survey. *J. Mach. Learn. Res.* **20**, 1–21 (2019).
32. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
33. Deng, J. et al. ImageNet: a large-scale hierarchical image database. in *IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
34. Foggia, P., Percannella, G., Soda, P. & Vento, M. Benchmarking HEp-2 cells classification methods. *IEEE Trans. Med. Imaging* **32**, 1878–1889 (2013).
35. Ulman, V. et al. An objective comparison of cell-tracking algorithms. *Nat. Methods* **14**, 1141–1152 (2017).
36. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
37. Sechidis, K., Tsoumakas, G. & Vlahavas, I. On the stratification of multi-label data. in *Machine Learning and Knowledge Discovery in Databases* Vol. 6913 (eds Gunopulos, D. et al.) 145–158 (Springer International Publishing, 2011).
38. Berman, M., Rannen Triki, A. & Blaschko, M. B. The Lovász-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. in *IEEE Conference on Computer Vision and Pattern Recognition* 4413–4421 (IEEE, 2018).
39. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? in *Advances in Neural Information Processing Systems* Vol. 27 (eds Ghahramani, Z. et al.) 3320–3328 (Curran Associates, Inc., 2014).
40. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. in *IEEE International Conference on Computer Vision* 618–626 (IEEE, 2017).
41. Deng, J., Guo, J., Xue, N. & Zafeiriou, S., Arcface: additive angular margin loss for deep face recognition. in *IEEE Conference on Computer Vision and Pattern Recognition* 4690–4699 (IEEE, 2019).
42. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
43. Ouyang, W., Mueller, F., Hjelmare, M., Lundberg, E. & Zimmer, C. ImJoy: an open-source computational platform for the deep learning era. https://doi.org/10.1038/s41592-019-0627-0 (2019).
44. Belthangady, C. & Royer, L. A. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* https://doi.org/10.1038/s41592-019-0458-z (2019).
45. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
46. Riley, P. Three pitfalls to avoid in machine learning. *Nature* **572**, 27–29 (2019).
47. Oei, R. W. et al. Convolutional neural network for cell classification using microscope images of intracellular actin networks. *PLoS ONE* **14**, e0213626 (2019).
48. Kornblith, S., Shlens, J. & Le, Q. V. Do better imagenet models transfer better? in *IEEE Conference on Computer Vision and Pattern Recognition* 2661–2671 (IEEE, 2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

Methods

Competition and prizes. This paper describes the outcome of the HPA Image Classification competition at Kaggle (https://www.kaggle.com/c/human-protein-atlas-image-classification/), which was active from 3 October, 2018 to 10 January, 2019. The top-ranking solutions were awarded with a cash prize (Table 1, first place, US$14,000; second place, US$10,000; third place, US$8,000 and fourth place, US$5,000).

Image generation in the HPA Cell Atlas. In the HPA Cell Atlas, each target protein is imaged in three different cell lines selected based on messenger RNA expression data from a large panel of cell lines. A standard immunostaining protocol is applied in a 96-well format to stain the target protein[49] (https://www.protocols.io/view/standardized-immunohistochemical-staining-used-in-yj8furw). To facilitate the downstream annotation process, reference markers for the nucleus, microtubules and endoplasmic reticulum are also stained. Confocal microscopes (63× oil immersion) are used to image each sample in the 96-well plate, and multiple images are typically acquired for each well. The four-color images (image size 2,048×2,048; 16-bit, pixel size 0.08 µm or 3,072×3,072; 16-bit, 0.08 µm) are uploaded to our laboratory information management system, where automatic quality checks are applied to pass only in-focus images with high contrast and good staining intensity. Manual annotation of the observed localization pattern is performed by applying one or more labels to the images (typically two to six) from each sample (same antibody, cell line and sample preparation date[3]). This dataset contains a mixture of images annotated by two types of workflow: (1) experts annotate the image, another expert curates it and (2) gamers from EVE Online[5] annotate the images and the annotations are curated by experts, as described in (1). Typically, at least two images per antibody per cell line are selected to create the dataset. In HPAv18, 32 labels are used to describe cellular localization classes.

Dataset assembly and quality control. The total dataset consisted of 42,774 images. The training set had 31,072 images, and the test set had 11,702 images. We provided all images both in high resolution (a mix of 2,048×2,048 and 3,072×3,072 pixels, TIFF 8-bit image files) and low resolution (512×512 pixels, PNG 8-bit grayscale files). Note that the PNG files were directly resized from TIFF images. The size inconsistency for the TIFF images were due to the variation of the actual size of acquired area from the sample originally, but they all shared the same pixel resolution. A total of 32 organelle classes from the HPA were merged into 28 classes for the competition (see Supplementary Table 1 for details of how the classes were merged and Supplementary Table 2 for the distribution of the classes within the training and test sets). Since the HPA Cell Atlas includes over 30 different cell lines, we sampled across 27 main cell lines, and have a roughly equal distribution across different cell lines. Sampling was done in groups consisting of each cell line and organelle combination to achieve this effect. Still, some groups were smaller than others, due to the imbalanced nature of the total HPA data, both regarding the class labels applied and the cell lines used in the experiment (see Supplementary Table 3 for the cell type image distribution in the training and test sets).

The test dataset images were collected first from nonpublished images generated within HPA. We excluded images from the same biological sample (with a specific protein and cell line) already represented in the test set from the training set to avoid information leakage. The training set images were collected from both public and nonpublic HPA Cell Atlas images.

For quality control, we applied further image analysis to get an acceptable quality of the images in the competition dataset. This allowed us to use nonpublished images from the HPA Cell Atlas, which are of mixed quality (high and low) compared to the high quality of the published images. The quality control was based on cell count and image contrast. The cell count was performed on the red (microtubules) channel images, mainly by applying a Gaussian filter and otsu thresholding with the scikit-image[50] library and by removing objects smaller than 8,000 pixels. We required a minimum of five cells per image and excluded cells touching the image borders. A minimum contrast check was applied to the green channel (protein of interest) with the 'is_low_contrast' method from scikit-image. We set 'fraction_threshold' to 0.2 and 'upper_percentile' to 99.99. Low contrast images were discarded.

Image-wise classification task. In this competition, the classification task is performed image-wise mostly because our current annotations are made at this level for each protein/antibody and cell line. We do not provide cell-wise labels, although it may yield better performance due to single-cell variations. However, we expect the impact to be minimal because only ~2% of proteins vary in their localization patterns between cells in images[3] and because these images are assigned the labels for all patterns observed in the image. Future improvements along this direction may be targeted to these proteins show single-cell variations.

Image resolutions and external data. We provide the entire image dataset in high and low resolutions, and we allowed the participants to explore the use of external data and pretrained models. We believe this is important, because it allowed the participants explore not only the model itself, but also different strategies to fetch external data, augment the data and take advantage of pretrained models, which had already been shown to be effective. We also believe it is important to obtain better performance under realistic constraints.

For the hidden variable problem raised from the metric learning model, prohibiting the use of external training data (and pretrained networks) may reduce the same type of risk. However, this may also prevent teams from achieving the desired outcome, since external data can improve performance, as we observed with HPAv18. A compromise may be to restrict the set of information allowed for training, such as only location labels from HPAv18 to avoid finding weak correlations with extra information such as antibody identifiers.

Data leakage disclosure and fix during the competition. During the competition, participants notified us that there was a data leakage issue after comparing the public HPA Cell Atlas Images (HPAv18) with the test set images using similarity analysis (for example, perceptual image hashing). We identified that 148 out of 11,702 images in the test set (including 'validation_public' and 'test_private') were mistakenly included. We also noticed that all the leaked images contain rare class labels, and many of the leaked images were not identical to images in HPAv18 but highly similar, for example by coming from a different focal plane.

Shortly after the leakage was identified, most of the leaked images in 'test_private' (the final evaluation test set to generate the ranking on the private leaderboard) were removed from scoring, and the rest of the leaked images were swapped with unleaked images with the same labels from 'validation_public' to keep rare classes in both test sets. All participants were notified of the leak and the fix. Most teams detected leaks in their code and excluded those leaked images from their own validation dataset.

Performance criteria. The F1 scores computed in Fig. 2 were computed from the following equation:

$$\text{Precision} = \frac{T_\text{p}}{T_\text{p} + F_\text{p}}$$

$$\text{Recall} = \frac{T_\text{p}}{T_\text{p} + F_\text{n}}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where T_p denotes the number of images that are true positive, and F_n, F_p denote the false negative and false positive, respectively.

The macro F1 score is computed from:

$$\text{macro F1} = \frac{1}{n}\left(\sum_{i=1}^{n} \text{F1}\right)$$

During the challenge, we used macro F1 to compute the score for public leaderboard rankings. Each team was allowed to select two final submissions at the end of the challenge. Macro F1 was used to compute the score on the private leaderboard (if there were two submissions for a team, we took the maximum score), and this was used as the main criterion to award the teams.

To analyze the score distributions of all teams, we took the one submission per team that gave the maximum score on the private leaderboard. For Fig. 2d we computed the score for all, single- and multi-localized classes with macro F1. Per-class F1 scores in Fig. 2c were computed with F1 for each class with no averaging. Scores shown in tables were rounded down.

Public and private leaderboard. During the competition, teams were scored and ranked on a public leaderboard. To prevent overfitting to the public leaderboard, a subset (29%) of the test set was used for calculating the leaderboard scores, while the remaining part (71%) of the test set was preserved for the final evaluation. This is important because participants tend to optimize toward higher scores on the public leaderboard at the risk of overfitting to the test data.

Since the participants ran their own model and only submitted the predicted labels for the test set, it was mandatory for the top four winning teams to submit their models for further inspection.

CAMs. We used Grad-CAM[40] to produce the CAMs shown in Fig. 3 and Supplementary Fig. 13 from a chosen reference convolutional layer of the network under investigation. We generated CAMs for the following models in Fig. 3: Model 1 is densenet121_1024 from Team 1, Model 2 is inception-v3 from Team 3, Model 3 is densenet121_standard_no_crop and the metric learning model from Team 1. The architecture of these three models are shown in Supplementary Fig. 12. For generating CAMs, the reference convolutional layer of these models was set to Block3, Mixed_7c, Layer3 and Layer3, respectively. The generated CAMs were resized and overlaid on top of the corresponding input image. These models are also described in Supplementary Notes: experiment 18 in Supplementary Table 5b, experiment 8 in Supplementary Table 5f, experiment 30 in Supplementary Table 5b and model 5 in Supplementary Table 5a.

Feature visualization with UMAP. For the feature visualization in Fig. 4, we used the 1,024-dimension feature from the 'fc' layer (as shown in Supplementary Fig. 12) of the densenet121_1024 model from Team 1. It was projected with UMAP to reduce the dimensionality from 1024 to 2. For generating the UMAP, the number of neighboring points, minimal point distance and number of components metric were set to 15, 0.1 and 2, respectively. The distance metric was set to Euclidean distance. We processed all the images in the training set and the entire HPAv18 dataset, and the generated 2D vectors were then plotted as a scatter plot. The data points are color coded are corresponding to their annotated location.

For the live HPA-UMAP (https://tinyurl.com/y6nhf5bo), we wrote the ImJoy plugin in Javascript and used Plotly.js to make the plot. Images shown when clicking the data points were pulled from the proteinatlas.org dynamically.

Metric learning model results evaluation. The score boost for the metric learning is mostly because of identification of 'batch effects' defined as images derived from different regions of the exact same sample (antibody and cell line combination), and not mainly from improved performance for rare classes. In total, 935 images in the 'test_private' set has one other image from the same sample in HPAv18 and the metric learning model detects 647 (69.2%) of them. We found 270 images in the 'test_private' set where the classification model made wrong predictions, but were successfully corrected by the metric learning model. Among these images, 261 have at least one image from the same sample in HPAv18 and 34 of them belong to rare classes (rare class defined as containing fewer than 1,000 images in the dataset). However, there are 71 images from the 'test_private' set (including five rare images) that were correctly predicted by the classification model, but replaced into wrong labels by the metric learning model.

Statistical analysis. The plotting and statistical analysis were performed with Python 3.6, NumPy, SciPy, scikit-learn, Pandas, seaborn and Matplotlib. To test the Macro F1 score difference between single label and multi-label images a two-sample Kolmogorov–Smirnov test was performed for significance testing. The test returns a two-tailed P value. The scores for single label images are significantly higher ($P < 1.08 \times 10^{-5}$) than for multi-label images for all groups. The test was done for each group of teams, 1–10: $n = 10$ teams ($P < 1.08 \times 10^{-5}$), 11–100: $n = 90$ teams ($P < 3.96 \times 10^{-51}$), 101–500: $n = 400$ teams ($P < 5.22 \times 10^{-197}$) and 501–2,137: $n = 1{,}637$ teams ($P < 4.01 \times 10^{-186}$).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset used for the HPA competition is available at: https://www.kaggle.com/c/human-protein-atlas-image-classification. The external dataset HPAv18 is publicly available on the HPA: https://v18.proteinatlas.org/. A script is provided for downloading the dataset is available at https://github.com/CellProfiling/HPA-competition.

Code availability

Source code used to produce the figures has been released under permissive licenses at https://github.com/CellProfiling/HPA-competition. A HPA classification competition model zoo is being built to offer downloads of the top models generated during the competition. The model zoo can be found at https://modelzoo.cellprofiling.org.

The source code for the ImJoy plugin HPA-UMAP can be found at https://github.com/imjoy-team/example-plugins.

References

49. Stadler, C., Skogs, M., Brismar, H., Uhlén, M. & Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteom.* **73**, 1067–1078 (2010).
50. Van Der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

Acknowledgements

We thank all the participants of the Human Protein Atlas Image Classification competition. We also acknowledge the staff at Kaggle for providing a competition platform that enabled this study and the competition prize sponsors Leica Microsystems and NVIDIA. The staff of the HPA program provided valuable contributions, such as data storage and management, and J. Fall helped with project administrative tasks. Funding was provided by the Knut and Alice Wallenberg Foundation (grant no. 2016.0204) and the Swedish Research Council (grant no. 2017–05327) to E.L.

Author contributions

E.L. conceived the study. C.F.W., M.H., D.P.S., H.X., C.K. and E.L. designed and implemented the competition together with the Kaggle team. W.O., C.F.W., M.H., L.Å., H.X., A.J.C., C.K. and E.L. performed data analysis and model investigation. S.D. contributed as the Kaggle competition winner and with intermediate analysis for the paper. S.D., Y.W., R.W., X.Z., X.C., K.L.T., D.B., A.K., S.F., D.Panchenko, P.J., C.J., X.Y., H.Z., Z.G., J.X., C.L., Y.G., J.L., R.W., B.T., C.H., K.H., D.Poplavskiy and S.M.G. contributed as participants of the nine selected teams from the Kaggle competition. W.O., A.J.C., C.F.W., M.H. and E.L. wrote the manuscript. W.O., L.Å., M.H., A.J.C. and C.F.W. made the figures. C.K. and S.D. revised the manuscript. E.L. supervised and administered the project and acquired funding. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0658-6.

Correspondence and requests for materials should be addressed to E.L.

Peer review information Rita Strack was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Emma Lundberg

Last updated by author(s): Oct 14, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used Leica Application Suite Advanced Fluorescence (LAS AF, 2.6.0 build 7266) to collect images used in this work. However the image collection (including experimental procedures and samples/reagents such as cell line authentication and antibody validation) is described in previous publications and is not part of this work.
Data analysis	For the competition dataset we mainly used Python 3 (with libraries such as numpy 0.17.0, pandas 0.24.2, scikit-image 0.15.0, scikit-learn 0.21.3, scipy 1.1.0, matplotlib 2.2.2, seaborn 0.9.0). Software libraries used by the 9 selected teams are described in their own source code available at https://modelzoo.cellprofiling.org

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The dataset used for the HPA competition is available at: https://www.kaggle.com/c/human-protein-atlas-image-classification

The external dataset HPAv18 is publicly available on the Human Protein Atlas: https://v18.proteinatlas.org/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	n/a
Replication	n/a
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

nature methods

Resource

https://doi.org/10.1038/s41592-024-02241-6

Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations

Received: 23 May 2023

Accepted: 11 March 2024

Published online: 9 April 2024

Check for updates

Srinivas Niranj Chandrasekaran[1], Beth A. Cimini[1], Amy Goodale[1], Lisa Miller[1], Maria Kost-Alimova[1], Nasim Jamali[1], John G. Doench[1], Briana Fritchman[1], Adam Skepner[1], Michelle Melanson[1], Alexandr A. Kalinin[1], John Arevalo[1], Marzieh Haghighi[1], Juan C. Caicedo[1], Daniel Kuhn[2], Desiree Hernandez[1], James Berstler[1], Hamdah Shafqat-Abbasi[1], David E. Root[1], Susanne E. Swalley[3], Sakshi Garg[2], Shantanu Singh[1] ✉ & Anne E. Carpenter[1] ✉

The identification of genetic and chemical perturbations with similar impacts on cell morphology can elucidate compounds' mechanisms of action or novel regulators of genetic pathways. Research on methods for identifying such similarities has lagged due to a lack of carefully designed and well-annotated image sets of cells treated with chemical and genetic perturbations. Here we create such a Resource dataset, CPJUMP1, in which each perturbed gene's product is a known target of at least two chemical compounds in the dataset. We systematically explore the directionality of correlations among perturbations that target the same protein encoded by a given gene, and we find that identifying matches between chemical and genetic perturbations is a challenging task. Our dataset and baseline analyses provide a benchmark for evaluating methods that measure perturbation similarities and impact, and more generally, learn effective representations of cellular state from microscopy images. Such advancements would accelerate the applications of image-based profiling of cellular states, such as uncovering drug mode of action or probing functional genomics.

Image-based profiling of cell samples is proving increasingly useful for biological discovery[1]. In image-based profiling, cells are treated with perturbations of interest and the resulting morphology is captured by microscopy. Cell morphology is quantitatively compared with identify meaningful similarities and differences among the perturbations, in the same way that transcriptional profiles are used to compare samples. More than a dozen applications have been demonstrated[1], including identification of the mechanisms of a disease by comparing cells from patients with a disease to those of healthy patients, identification of the impact of a chemical compound by comparing cells treated with it to untreated cells, and identification of gene functions by clustering large sets of genetically perturbed samples to determine relationships among the genes.

Typically, morphological features are extracted from each cell using classical image processing software. These so-called 'hand-engineered' features have been carefully developed and optimized to capture cellular morphology variations, including size, shape, intensity and texture of the various stains in the image. These features are the current standard in the field and require post-processing steps including normalization, feature selection and dimensionality

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Merck Healthcare KGaA, Darmstadt, Germany. [3]Biogen Inc., Cambridge, MA, USA.
✉e-mail: shantanu@broadinstitute.org; anne@broadinstitute.org

https://doi.org/10.1038/s41592-024-02241-6

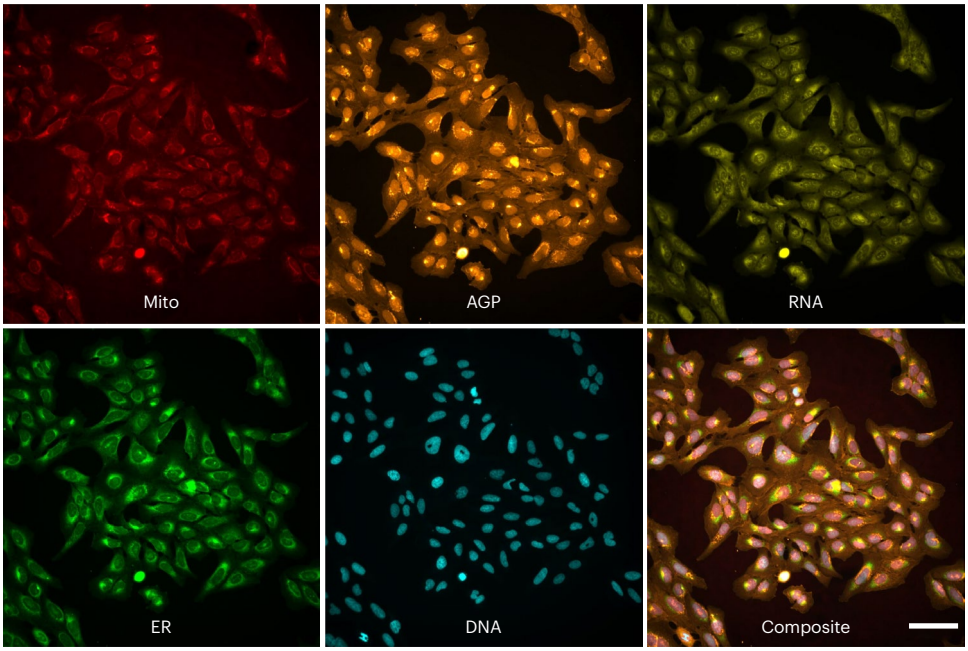


Fig. 1 | Sample images from the dataset. A five-channel image of human U2OS cells treated with the compound PFI-1 (a BRD4-specific inhibitor). This is a representative image from one of four wells of cells treated with PFI-1. The channel names indicate the cellular structures identified in each image (see Methods section for details; AGP, actin, Golgi, plasma membrane; DNA, nucleus; ER, endoplasmic reticulum; mito, mitochondria; RNA, nucleoli and cytoplasmic RNA). Other example images (including brightfield channels not shown here) are available at https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/tree/6ba3fcd1495d9e844e4607373a568641981ffcd8/example_images. Scale bar, 100 µm.

reduction[2]. With advances in representation learning during the last decade, it is natural to ask what set of features could be automatically identified by machine learning algorithms, directly from pixels.

However, image-based profiling has yet to fully benefit from the latest machine learning research. The vast majority of studies use classical segmentation and feature extraction; deep learning methods are beginning to be explored and there is much room for advancement[3,4]. Historically, the lack of ground truth has been a major limiting factor in the field; that is, the true relationships among perturbations (for example, genes and compounds) are unknown and require significant effort to ascertain[5]. Although this is exciting because the potential for biological discovery is high, the lack of ground truth presents a challenge for optimizing deep learning pipelines.

Here we describe our design and creation of a benchmark dataset via a single large experiment, CPJUMP1. The dataset consists of approximately 3 million images of cells, image-based profiles of 75 million single cells, and well-level aggregated profiles. A sample five-channel image is shown in Fig. 1. This dataset contains chemical and genetic perturbation pairs that target the same genes and are tested in separate wells to see whether they produce similar (or opposite) phenotypes. Although these pairings are not absolute truth for a number of reasons discussed later, they are nevertheless more likely than random pairs to match (that is, induce similar or opposite morphology changes). This Resource is unique because there are no other image-based datasets of the Cell Painting assay (described later) that include pairs of annotated genetic and chemical perturbations performed side by side, under different experimental conditions such as different cell types, time points and imaging conditions. These were also executed in parallel to minimize technical variations that may confound the signal. There are many public Cell Painting datasets (for example, https://github.com/broadinstitute/cellpainting-gallery) but we are aware only of one with genetic and chemical perturbation types run in parallel (RxRx3, https://www.rxrx.ai/rxrx3). RxRx3 has not been provided with gene–compound relationship annotations and all but 733 genes are anonymized; any pairs that exist would be scattered across many plates and batches. As well, it includes only a single cell type, time point and imaging condition. Thus, the nature of CPJUMP1 enables the testing of computational strategies to optimally represent the samples so that they can be compared and thus uncover valuable biological relationships. It also enables comparison of CRISPR-Cas9 knockout and ORF (open reading frame) overexpression as mechanisms to perturb genetic pathways and to identify the compounds' mechanisms of action.

Results

To push forward advancements in this field, we assembled a consortium of 10 pharmaceutical companies, two non-profit institutions, and several supporting companies, known as the JUMP Cell Painting Consortium (Joint Undertaking in Morphological Profiling). Members of this Consortium created the ground truth dataset we present here, for optimizing the main assay used in image-based profiling, called the Cell Painting assay, and to move methods in the field forward[6,7]. We selected and curated a set of 160 genes and 303 compounds with (relatively) known relationships between each other, and designed an experimental layout to enable testing and comparing methods to quantify their similarities (Methods), all with a strong emphasis on making this dataset useful for developing computational methods for the field.

There are two groups of experimental conditions in this dataset, the primary group and the secondary group. In the primary group we separately captured chemical and genetic perturbation (CRISPR knockout and ORF overexpression) profiles in two cell types (U2OS and A549) at two time points (Supplementary Table 1; a representation of profiles from a subset of this experiment is shown in Fig. 2). There are a total of 40 384-well plates in the primary group of experimental conditions (Fig. 3). The secondary group consists of additional plates of experimental conditions as well as plates from the primary group that have undergone additional imaging conditions (Fig. 3), which are described in the Methods section. In addition to being used to optimize the assay conditions[7], the primary and secondary groups offer multiple views of cells treated with each chemical or genetic perturbation, and therefore can be used for many interesting machine

https://doi.org/10.1038/s41592-024-02241-6

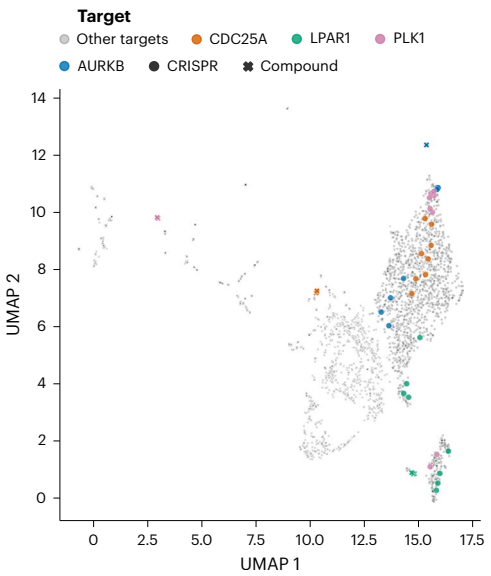


Fig. 2 | UMAP representation of image-based profiles from a subset of the primary group of tested samples. Profiles from human A549 cells perturbed by all compounds and CRISPR guides at the long time points (Supplementary Table 1) are shown here. The top four most-similar compound–genetic perturbation pairs, associated with the same gene, are highlighted.

learning explorations, such as style transfer (for example, to attempt prediction of one experimental condition from another), information retrieval and multi-view learning, and for benchmarking representation learning methods. There are 11 additional plates in the secondary group, but there are 67 plates of images because several plates were imaged multiple times (Fig. 3; Methods).

A major goal of image-based profiling is to derive a representation from the quantitative images of cell samples, such that samples in biologically similar states have similar representations. Given such a representation, solutions for many of the applications discussed become immediately feasible. Here we benchmark representation learning as a foundation for methods development in the future.

As a way to compare different representation methods, we created benchmarks based on two tasks: detection of differences between perturbations and negative controls to identify active perturbations, and the grouping of gene–compound pairs in which the gene's product (protein) is a target of the compound (as well as grouping two CRISPR guides targeting the same gene, or two compounds annotated with the same target). For both tasks we use cosine similarity (or its absolute value), a simple but widely used correlation-like metric, to measure similarities between pairs of well-level aggregated profiles. In some cases the expected directionality of correlation is positive while in other tasks the correlations may be strongly positive or negative; we adjust statistical tests for each task accordingly.

Benchmarking perturbation detection methods

We chose perturbation detection as one of the tasks to evaluate representations because it often precedes other useful applications (by removing samples that have no or little true signal), and is equivalent to measuring statistical significance of the perturbation's signal. For example, a set of chemical or genetic perturbations might be filtered by this phenotypic activity criterion before embarking on subsequent laboratory experiments, or prior to training a model, or other analysis that could be confounded by noisy signals. It can also be useful for determining which experimental protocol or computational analysis pipeline is most sensitive from among several alternatives. It should be noted that even given perfect computational methods for feature extraction, batch correction and profile comparison, many samples will be detectably different from negative controls for several biological reasons. For example, a chemical or genetic perturbation may affect cell morphology only in a particular cell type, under particular environmental conditions, at a particular time, or if particular stains were used, conditions that may not have been met in the experiment. Conversely, a perturbation's impact may be amplified by the plate layout, given that even unrelated perturbations in the same well position might look similar. This concern is overcome by matching treatments in different well positions, where such data are available (see Benchmarking perturbation matching methods).

We used average precision to measure each primary group sample's ability to retrieve its replicates against the background of negative control samples, using cosine similarity as the similarity metric. The significance of the average precision value is assessed using permutation testing to obtain a P value, which is then adjusted using the false discovery rate to yield a corrected P value (q value). We calculate the mean of the average precision for each perturbation and then term the fraction of perturbations with a q value below the significance threshold (0.05) as the fraction retrieved. Details about the computation of average precision and fraction retrieved are provided in the Methods section.

In general, we find that the fraction retrieved for compounds is higher than that of genetic perturbations, across all conditions (Fig. 4a). This indicates that chemical compounds produce phenotypes that are more distinguishable from negative controls, compared with phenotypes produced by CRISPR knockout and ORF overexpression. We also find that the fraction retrieved is higher for CRISPR knockout than for ORF overexpression (Fig. 4a). In summary, compounds, CRISPRs and ORFs all yield signals in the assay, with the compounds being the strongest and ORFs the weakest. However, we emphasize a strong technical variable that precludes a strong conclusion here: the reduced fraction retrieved values for ORF may be attributed to plate layout effects, in which identical treatments in different rows or columns have dissimilar profiles. This factor amplifies the systematic technical noise in the compound and CRISPR plates due to their particular layout, while it adversely impacts ORFs (Methods). Retrieving the same position replicates for ORF does increase the fraction retrieved, as would be expected if plate layout effects are substantial (Extended Data Fig. 1). Plate layout effects may be partly mitigated by mean centering every feature at each well position, although we have not applied the correction to this dataset because we do not have sufficient diversity of samples in each well position across a large number of plates. Furthermore, although there is a presence of signal for all three perturbation modalities, we note that these were not random sets of genes and compounds; instead, the compounds were chosen from the Drug Repurposing set[8] and the genes were selected secondarily depending on whether their protein products are targeted by those compounds. This is likely to select for reagents that yield phenotypes versus random genes and compounds.

Benchmarking perturbation matching methods

We next established a benchmark for researchers to develop and test strategies for a real-world retrieval task, in which we search for genes or compounds that have a similar impact on cell morphologies as the query gene or compound. Improved methods would enable improved discovery of the compounds' mechanisms of action based on a compound query[9], and virtual screening for useful compounds based on a gene query[10]. This dataset presents a unique opportunity to match profiles of perturbations across modalities (chemical versus genetic) because genes in this dataset are targeted by two types of genetic perturbations (ORF and CRISPR-Cas9 knockout) and by at least two compounds. Similarly, because there are both a pair of CRISPR guides and a pair of compounds targeting each gene or gene's product, this dataset can be used to match profiles within a perturbation modality (there is only one ORF reagent per gene, therefore a similar analysis is

https://doi.org/10.1038/s41592-024-02241-6

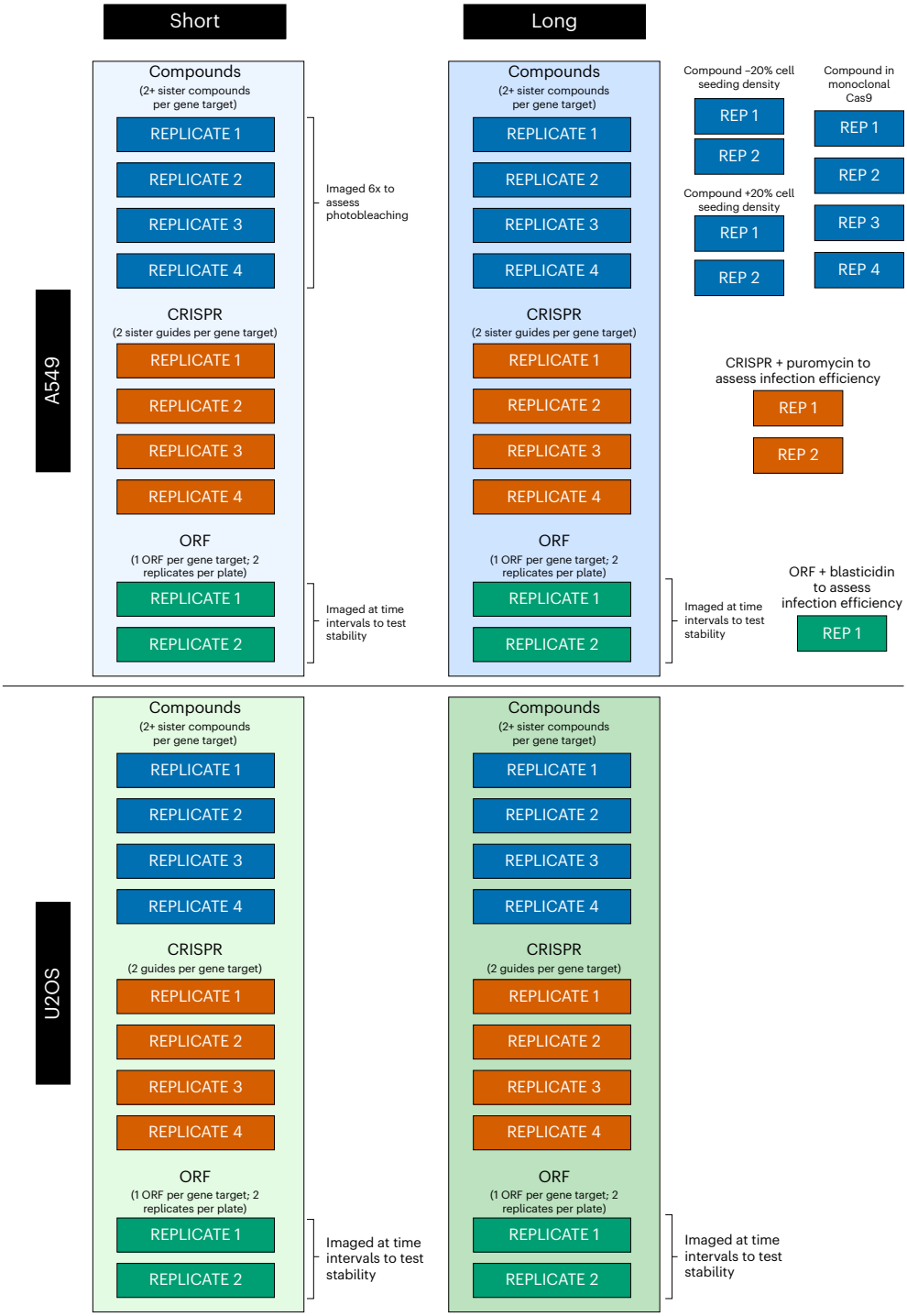


Fig. 3 | Schema of the data generated in the CPJUMP1 experiment. Each rectangular block of replicates (REPLICATE X or REP X) is a 384-well plate of cells perturbed by compounds, CRISPR guides or ORFs and subjected to different experimental conditions. Short and long time points are described in Supplementary Table 1. Plates in the vertical blue and green boxes comprise the primary CPJUMP1 experiment. The other (that is, secondary) experimental conditions are described in the Methods section.

not possible for overexpression). It also offers an opportunity to study the directionality of profile matching; for example, whether CRISPR knockouts and ORF overexpressions consistently yield anti-correlated profiles.

After filtering out perturbations that were indistinguishable from negative controls ($q > 0.05$), we then evaluated average precision to identify true connections (that is, perturbation pairs that target the same gene or gene's product), which are distinct from false connections (that is, pairs not known to target the same gene or gene's product).

We first tested the ability to retrieve true connections within the same perturbation modality: that is, 'sister' compounds that are annotated as targeting the same gene should match each other, and 'sister' CRISPR guides that target the same gene should match each other. Because compounds can enhance or inhibit the function of a protein (or have other impacts), those with the same gene annotation might be positively correlated or negatively correlated; for compounds, we therefore used the absolute value of cosine similarities while calculating average precision. By contrast, for CRISPR guides, those targeting

https://doi.org/10.1038/s41592-024-02241-6

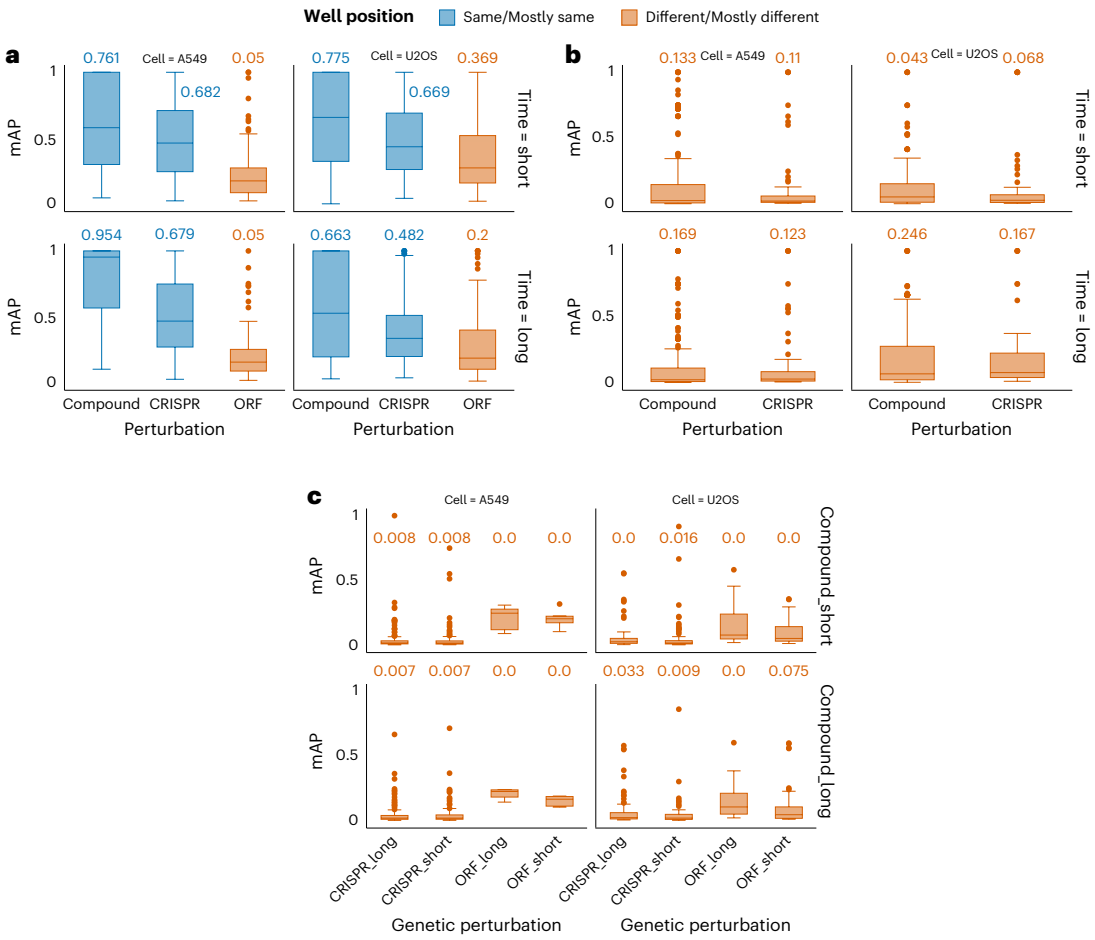


Fig. 4 | Benchmark results for progressively more difficult retrieval tasks. **a**, Perturbation detection: retrieving replicates of the same sample. Mean average precision (mAP) for perturbation detection is shown across experimental conditions: cell type (columns) and time points (rows; short and long time points are defined in Supplementary Table 1). The numerical values shown above each box plot are the fraction of perturbations that can be successfully retrieved for each retrieval task. Box plot boundaries are 75th (Q3) and 25th (Q1) percentiles, with whiskers at ±1.5-fold the interquartile range (Q3–Q1) or the highest or lowest data point. The color of the bars denotes whether the query perturbation and the retrieved perturbation are in mostly the same well position (blue) or mostly different well positions (red); the latter is a more challenging task due to technical well-position artifacts. **b**, Perturbation matching, within a perturbation type: the plot shows mAP for sister perturbation retrieval (that is, pairs of compounds or pairs of CRISPR guides annotated with the same gene target). ORFs are not shown because there is only a single ORF reagent per gene. Absolute cosine similarity is used for calculating mAP values for compounds because pairs of compounds annotated to target the same protein can be positively or negatively correlated. **c**, Perturbation matching, across perturbation types: the plot shows mAP values for retrieving compound–gene pairs (that is, the same target and different perturbation type). Absolute cosine similarity is used for calculating mAP values for both compound–CRISPR and compound–ORF matching. The number of independent biological samples is available in Supplementary Table 2.

the same gene are expected to be positively correlated, therefore we used the actual values of cosine similarities, as in the perturbation detection task.

With baseline methods, this task is (not surprisingly) much more challenging than retrieving replicates of the same sample (compare Fig. 4a,b). Of all of the compounds that yield a signal, only ~5–25% of them correctly match their sister compounds targeting the same protein. Likewise, and more surprisingly given their expected accurate annotation and specificity, only 7–17% of the CRISPR reagents correctly match to their sister guides targeting the same gene. We cannot distinguish the many factors that are likely to make this a challenging task, including non-optimal ground truth annotations for compounds, off-target effects (for compounds and CRISPR guides), differing levels of knockdown for CRISPR guides, lack of information content in the assay, polypharmacology in which each compound impacts multiple targets (see Discussion) and/or non-optimal methods for matching samples. Although the values of fraction retrieved are similar for compounds and CRISPR guides, more compounds are distinguishable from negative control than CRISPR guides (Fig. 4a and Supplementary Table 2). Thus, surprisingly, retrieving sister compounds is more successful than retrieving sister CRISPR guides (Fig. 4b), perhaps because compounds tend to induce stronger phenotypes (Fig. 4a).

Next, we assessed cross-modality matching: that is, the ability to retrieve correct gene–compound pairs. Retrieving compound–gene pairs is more difficult than perturbation detection and sister perturbation retrieval (which itself reached only ~25% in the best scenario), but it is extraordinarily useful for identifying novel chemical regulators of genes and identifying the mechanism of query compounds. Even a low success rate, therefore, can accelerate drug discovery. Given the potential for compound–gene matches to be positive or negatively correlated (detailed in the next paragraph), we used the absolute cosine similarities for both compound–ORF and compound–CRISPR retrieval.

Gene–compound retrieval is only slightly better than expected by chance (Fig. 4c) and, as expected, is less effective than compound–compound and CRISPR–CRISPR matching (Fig. 4b). This might reflect that gene–compound annotations are less reliable than compound–compound annotations (and certainly less reliable than CRISPR guide annotations, for which the target gene is designed a priori to be accurate and specific), and/or that better methods are needed to align data across modalities. Comparing the two genetic perturbation

https://doi.org/10.1038/s41592-024-02241-6

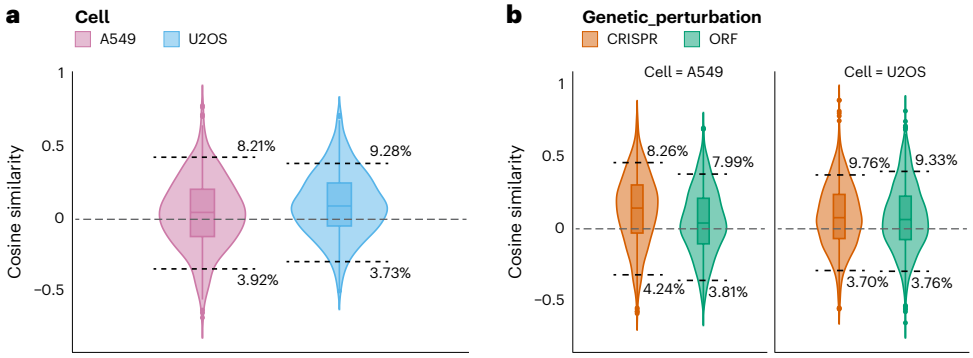


Fig. 5 | Directionality of matching cross-modality. a, Cosine similarity between ORF and CRISPR reagents targeting the same gene is shown for the two cell types, A549 (pink) and U2OS (blue). The 5th and 95th percentile thresholds of their respective nulls (ORFs and CRISPR reagents targeting different genes), along with the percentage of true pairs below the thresholds, are shown. We performed Fisher's exact test to determine whether the true pairs beyond the threshold are more likely to be positively correlated or negatively correlated. We find them to be significantly more positively correlated (*P* values are available in Supplementary Table 3). **b**, Cosine similarity between compounds and the two genetic perturbation modalities, CRISPR (orange) and ORF (green) targeting the same gene or gene product. All analyses here were also statistically significantly more positively correlated; *P* values are available in Supplementary Table 4. Cosine similarity of zero is shown as a gray dashed line in both subplots. In **a** there were $n = 3{,}728$ biological independent ORF and CRISPR reagents and in **b** there were $n = 1{,}864$ independent pairs of compounds and genetic perturbations targeting the same gene or gene product. Box plot boundaries are 75th (Q3) and 25th (Q1) percentiles, with whiskers at ±1.5-fold the interquartile range (Q3–Q1) or the highest or lowest data point.

modalities, we find that the values of mean average precision for retrieving compound–CRISPR pairs were better than that of compound–ORF pairs across various cell types and time points, except at one time point, where ORFs match better to compounds than CRISPR reagents do (Fig. 4c).

Although the performance for compound–genetic perturbation retrieval is low compared with the other retrieval tasks discussed above, it should be strongly noted that significant time and resources are otherwise required to identify the target of a compound, and similarly to identify compounds that target a particular gene. Therefore, even the baseline's relatively low matching rates might accelerate drug development by yielding a list of possibilities for biologists to test directly in subsequent experiments, or to be combined with orthogonal lists of candidate targets for a compound, to improve accuracy. Improving image representations and thus the accuracy of predicted matches by a few percentage points could therefore have a major impact on the discovery of compounds that impact proteins of interest, and the identification of the mechanism of action of compounds of interest.

Finally, we examined the directionality of gene–compound matching, that is, whether a compound targeting a protein encoded by a given gene has a correlating or anti-correlating profile with CRISPR (which reduces the amount of gene product) and with ORF (which increases the amount of gene product). Most compounds are annotated as inhibiting the function of their target gene's product, therefore one might expect image-based profiles from cells treated with CRISPR guides to generally positively correlate to (mimic) the corresponding compound's profile, whereas ORF profiles might generally be expected to anti-correlate (oppose) the corresponding small molecule's profile because overexpression often increases a gene's function. By the same rationale, ORFs and CRISPR guides targeting the same gene might be expected to yield opposite (anti-correlated) effects on the cells' profiles. However, we strongly note that there will be numerous exceptions, given the nonlinear behavior of many biological systems, and any number of distinct mechanisms for which these general principles may not hold, which we have previously detailed[10]. For example, many compounds do not inhibit their target protein's function but instead activate it or induce some new function, and many overexpressed genes may have no impact at all, or even have a dominant negative or feedback loop or compensatory impact on the gene's function. Furthermore, the choice of cell type, time point or readout for capturing the similarity may not be optimal. In fact, the exceptions may be more common than the commonsense rules in this case. One aim of generating this dataset was to quantify how often the expected relationships and directionalities occur, to provide concrete evidence in this so-far theoretical debate with only anecdotal evidence available.

We began by testing the basic hypothesis that CRISPR and ORF reagents targeting the same gene should yield negatively correlated (opposite) profiles to each other. Surprisingly, we found that the CRISPR and ORF profiles are slightly positively correlated with each other, in both cell types (Fig. 5a and Supplementary Table 3). We then compared the cosine similarities between compound–CRISPR pairs and compound–ORF pairs that target the same gene. In both U2OS and A549 cells, we found that both compound–CRISPR and compound–ORF pairs were more positively correlated than negatively correlated (Fig. 5b and Supplementary Table 4).

Looking at the strongest-matching positive and negative gene–compound pairings (Supplementary Tables 5 and 6), we found many pairings with explainable directionality (for example, CRISPR knockout of a gene matches a compound annotated as inhibiting the protein product of that gene). For example, the top positively correlated gene–compound match in U2OS cells is the PLK1 inhibitor compound BI-2536 matched with CRISPR against PLK1 (Extended Data Fig. 2), and the next two matches are annotated as Aurora kinase inhibitors that match CRISPR against AURKB (Supplementary Table 5). Similarly, the five strongest matches in A549 cells (Supplementary Table 6) are all CRISPR reagents positively correlating with compounds annotated as targeting the correct protein encoded by a given gene. Some overexpressions also match the expected directionality, such as the second strongest negatively correlated match in U2OS cells, which is the compound GSK2110183, an AKT inhibitor negatively correlated to overexpression of AKT1. Still, there were many pairings with unexpected directionality, possibly due to ORFs exhibiting dominant-negative behavior, as in the seventh positively correlated gene–compound match in U2OS cells, the compound pentostatin, which is annotated as an adenosine deaminase inhibitor; it matches with overexpression of ADA (adenosine deaminase). Unexpected directionality is evident when examining BRD4 overexpression and BI-2536 treatment, both of which induce cell death (Extended Data Fig. 2). Because BI-2536 inhibits BRD4, one might expect BRD4 knockout to closely align with BI-2536's effects. This discrepancy suggests two possibilities: either the multi-target nature of BI-2536 leads to a dominant phenotype (PLK1 inhibition may be stronger than BRD4 inhibition), or BRD4 inhibition

has a limited phenotypic impact. This latter possibility aligns with the negative control-like phenotype observed with the BRD4-specific inhibitor, PFI-1 (Extended Data Fig. 2). Some CRISPR reagents produce surprising directionality, as in the case of the compound TG-003 in U2OS cells, annotated as a CLK inhibitor but negatively correlated to CRISPR knockout of CLK1. Given these findings, and the possibility that compounds often behave differently in different cellular contexts and may be annotated based on a particular one, it is not surprising that compounds targeting the products of particular genes do not show a consistent directionality relationship with ORFs or CRISPRs.

Discussion

Biology would benefit greatly from the machine learning community turning its attention to rich, single-cell imaging data. Although our results may only touch upon the potential applications of machine learning, our emphasis is a strategic appeal to the machine learning community. We hope that the Resource and benchmarks we have created will provide a foundation on which researchers can develop and test novel methods for representation learning, multi-view learning, information retrieval and style transfer, among others. The task of identifying targets of a compound to understand its mechanism is exceptionally difficult, expensive and time-consuming, creating a major bottleneck in developing useful compounds for chemical biology and drug discovery[11]. This has several implications for this work: first, even with low but non-zero success rates, biologists can use their understanding of the compound's known traits to create and test hypotheses about a chemical's mechanism of action, accelerating discovery. Second, the predictions from this method might be combined with other predictive approaches, such as rank ordered candidate gene lists from structure-based chemical–protein binding predictions, to produce more reliable results. And last, we also emphasize that, as a direct result of the fact that identifying the targets of compounds is so difficult, only a very limited amount of rather noisy ground truth exists; each known compound–gene interaction has been painstakingly discovered after hundreds of thousands of dollars of effort over many years, and many pairings are uncertain. By contrast, many mainstream machine learning tasks are oriented to replicate specific human skills where ground truth can be collected at large scale (for example, translation or image recognition), and, given sufficient resources, the accuracy often approaches 100%. We believe that supervised methods hold promise[1], and we hope that novel machine learning methods developed using our dataset will be used to discover new gene–compound connections that can add to the ground truth for this problem in the future.

Beyond biology, our dataset provides a challenging, real-world test bed for many kinds of more general machine learning algorithms. It is a large-scale perturbation experiment with complex multidimensional, hierarchical data (images displaying dozens of cells each), and we believe that new machine learning strategies still need to be developed to realize its full potential. In addition to the prediction problems we present in this paper, it also opens up problems in high-level reasoning on experimental data, enabling the study of complex artificial intelligence strategies, such as causal inference (observations from interventional experiments), planning (optimizing the next intervention that maximizes discovery), and simulations (what would have happened if other interventions are applied).

Finally, unusual aspects of the data type that we present pose challenges to machine learning algorithms and will require that they be pushed in different directions to adapt. This may spark creative solutions with broader impact in machine learning. For example, multiplexed imaging will push the field of machine learning to adapt to domains outside the red–green–blue (RGB) colorspace of natural images in which the number and relationship between the channels (for example, the extent of correlation) is very different, compared with natural images of everyday objects.

In general, deep learning-based features may provide improvements in performance for some tasks over the classical CellProfiler-based features, although interpretability remains a challenge. The direct interpretability of CellProfiler-based features can help in examining signatures and comparing them to evidence found in existing research relevant to a particular profiling task. One can readily visualize any target cell population by looking up the associated sample cell images or by converting CellProfiler representations into images with convolutional neural network-based image generators. An example set of cell images generated by a fundamental version of such a model is shown in Extended Data Fig. 3. The research community can delve deeper into creating interpretative models from CellProfiler profiles that we have provided in this resource.

Still, this dataset has limitations. It covers only 160 genes and 303 compounds, and ~21% of compounds in this dataset are annotated as targeting proteins in more than one gene family (Supplementary Table 7). Ideally, such a dataset would include only compounds that are very well-studied as targeting one and only one gene's product. In reality, this is impractical. Polypharmacology is increasingly recognized as common for compounds[12,13], and this is likely to substantially impact the ability of compounds' images to match one of the annotated genes' images. As well, the target annotations of the compounds may not be complete because of undiscovered gene–compound relationships.

We also note limitations in the selection of genes and compounds. This dataset was curated with compounds and genes available in the Broad Institute's Drug Repurposing Hub and in its library of genetic perturbation reagents, respectively, which introduces several biases. First, the set contains only preclinical or clinical compounds with stronger binding and higher specificity than randomly synthesized compounds; however, for the tasks of mechanism of action determination and virtual screening (in which the goal is ultimately to identify such compounds), this is not an overly concerning bias. Second, because of our selection criterion that at least two compounds should target the product of every gene, all compound–gene pairs without a second compound in the Drug Repurposing Hub were excluded, making better-studied compounds more represented. There were also other selection criteria that introduce various biases, such as that the selected compounds should not be a controlled substance. In terms of experimental conditions, we used a single concentration (5 µM) for all compounds, which is not ideal, and we created this dataset in a single experiment at a single facility; this choice minimizes technical variability and therefore maximizes the biological signal in the data, but also limits the potential for generalizability of any models using it as training data. We note that generalizability of models across datasets is often unnecessary in biology experiments, where controls can be included in each experiment; in fact, we recommend that those creating large datasets include the sets of compounds and genes we present here in the experiment to have internal controls and/or landmarks for the assay. Our consortium has adopted this approach in creating our large-scale dataset of 136,000 chemical and genetic perturbations[14].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02241-6.

References

1. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
2. Caicedo, J. C. et al. Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).

3. Pratapa, A., Doron, M. & Caicedo, J. C. Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.* **65**, 9–17 (2021).
4. Moshkov, N. et al. Learning representations for image-based profiling of perturbations. https://doi.org/10.1038/s41467-024-45999-1 (2022).
5. Canham, S. M. et al. Systematic chemogenetic library assembly. *Cell Chem. Biol.* **27**, 1124–1129 (2020).
6. Bray, M.-A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
7. Cimini, B. A. et al. Optimizing the Cell Painting assay for image-based profiling. *Nat. Protoc.* **18**, 1981–2013 (2023).
8. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
9. Perlman, Z. E. et al. Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
10. Rohban, M. H. et al. Virtual screening for small-molecule pathway regulators by image-profile matching. *Cell Syst.* **13**, 724–736 (2022).
11. Lill, J. R., Mathews, W. R., Rose, C. M. & Schirle, M. Proteomics in the pharmaceutical and biotechnology industry: a look to the next decade. *Expert Rev. Proteomics* **18**, 503–526 (2021).
12. Rastelli, G. & Pinzi, L. Computational polypharmacology comes of age. *Front. Pharmacol.* **6**, 157 (2015).
13. Proschak, E., Stark, H. & Merk, D. Polypharmacology by design: a medicinal chemist's perspective on multitargeting compounds. *J. Med. Chem.* **62**, 420–444 (2019).
14. Chandrasekaran, S. N. et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. Preprint at https://doi.org/10.1101/2023.03.23.534023 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

Methods

Compound and gene selection

The CPJUMP1 dataset consists of images and profiles of cells that were perturbed separately by chemical and genetic perturbations, in which both sets were chosen based on expected matching relationships between them. Chemical perturbations refer to small molecules (compounds) that affect the function of cells, while the genetic perturbations refer either to ORFs that overexpress genes (that is, yield more of the gene's product in the cell) or to guide RNAs that mediate CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats), which cause knockout of gene function (by decreasing production of the gene's product in the cell; the term 'knockout' is used in the field, although in the timescale of these experiments residual protein probably remains, depending on natural rates of protein turnover).

We therefore designed CPJUMP1 such that for each gene, we have one ORF, two CRISPR guides (for all but 15 genes), and one or two compounds that are thought to affect the cell by influencing the function of that gene's product (although they may also influence the function of other genes due to polypharmacology, complicating the signal; Supplementary Tables 8 and 9).

We derived the list of compounds from the Broad Institute's Drug Repurposing Hub dataset[8], a curated and annotated collection of Food and Drug Administration-approved drugs, clinical trial drugs and preclinical tool compounds (Extended Data Fig. 4d). The genes perturbed by genetic perturbations were chosen because their associated proteins are annotated targets of the compounds. The specific criteria for compounds, genetic reagents (considering their on- and off-target effects), and controls are described in the section 'Compound selection criteria', and their layout on the plates is described in the section 'Plate layout design'. After applying the filters and including controls, we selected a total of 303 compounds and 160 genes such that their corresponding perturbations could fit into three 384-well plates with controls.

Compound selection criteria

We filtered the Drug Repurposing Hub compounds using several criteria, of which three are most important:

- The compounds should target proteins encoded by genes that belong to diverse gene families (Supplementary Table 10). This is because methods for representation learning and gene–compound matching should work well for many different biological pathways, rather than for only a few that are well-characterized and/or easy to predict.
- Each gene product should be targeted by at least two compounds, so that gene–compound matching and compound–compound matching can both be performed using the dataset.
- The requirement that each compound should target only a single gene product, which was considered, but which is difficult to achieve due to polypharmacology (Supplementary Tables 8 and 9), that is, the property for compounds to bind and affect many different gene products in the cell; this is especially common for protein kinase inhibitors in the dataset. Instead, we filtered out only the so-called 'historical compounds' listed in the Chemical Probes Portal[15], which are compounds that are known to be quite non-selective (or not sufficiently potent) compared with other available chemical probes.

Our list of compounds and genes also includes both negative and positive controls. The negative controls for each perturbation modality are, first, the compounds (DMSO, that is, dimethyl sulfoxide, which is the solvent for all of the compounds studied; in other words, all samples will have DMSO added at the same concentration, but the negative controls have no additional compound added); second, ORFs (15 ORFs with the weakest signature in previous image-based profiling experiments[16]; thus, the total number of genes with ORFs is 160 + 15 = 175); and third, CRISPR guides (30 CRISPR guides that target an intergenic site (cutting controls, $n = 3$) or do not have a target sequence that exists in human cells (non-cutting controls, n = 27)).

There are three types of compound positive controls in our list. First, we included chemical probes that are very well-studied and (unlike most compounds) are known to very selectively modulate the target genes' product (poscon_cp)[15]. Second, we included compounds that strongly correlate with the correct genetic perturbation in previous image-based profiling experiments with ORFs[16] and compounds (poscon_orf)[17]. Finally, we included a set of very diverse pairs of compounds with strong intra-pair and weak inter-pair correlations, based on prior experiments (poscon_diverse).

Additionally, compounds were filtered based on availability from at least one of four compound vendors (Sigma, SelleckChem, Tocris and MedChemEx) and genetic reagents via the Broad Insitute's Genetic Perturbation Platform portal. Last, we also excluded compounds on the U.S. Drug Enforcement Agency (DEA) list of controlled substances or the Organisation for the Prohibition of Chemical Weapons (OPCW) list of chemical weapons precursors.

Target loci selection

We picked the target loci for the CRISPR experiments by selecting the top-two-ranking single-guide RNA sequences that maximize their on-target activity, calculated using the Azimuth 2.0 model[18], and minimize the off-target activity, calculated using the Cutting Frequency Determination score (additional details can be found at https://portals.broadinstitute.org/gpp/public/software/sgrna-scoring-help).

Compound and gene metadata

A list of CRISPR reagents and their target sequences is available here: https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/blob/6ba3fcd1495d9e844e4607373a568641981ffcd8/metadata/external_metadata/JUMP-Target-1_crispr_metadata.tsv

A list of ORF reagents and their target sequences is available here: https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/blob/6ba3fcd1495d9e844e4607373a568641981ffcd8/metadata/external_metadata/JUMP-Target-1_orf_metadata_with_sequence.tsv

A list of compounds with their names, PubChem unique identifier, SMILES (Simplified Molecular Input Line Entry System) string, and gene product targets is available here: https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/blob/6ba3fcd1495d9e844e4607373a568641981ffcd8/metadata/external_metadata/JUMP-Target-1_compound_metadata_targets.tsv

Infection efficiency data for the ORF and CRISPR regents for each time point and cell type from the CellTiter-Glo cell viability assay are available here: https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/blob/6ba3fcd1495d9e844e4607373a568641981ffcd8/metadata/external_metadata/CPJUMP1_Infection_Efficiency.xlsx

Plate layout design

The first plate (Extended Data Fig. 4a) consists entirely of compounds, with two compounds per gene product target. Each compound is in singlicate on the plate except for a dozen or so compounds (poscon_diverse) in duplicate and the negative control DMSO described above, in $n = 64$ replicates. The second plate (Extended Data Fig. 4b) consists entirely of CRISPR reagents, with two guides per gene, each arrayed in its own well and kept separate, with no within-plate replicates; there are two replicates of the 30 CRISPR negative controls described above. The third plate (Extended Data Fig. 4c) consists entirely of ORFs: because there was only one perturbation reagent per gene, there are two replicates of each per plate, plus $n = 4$ replicates of the 15 ORF negative controls. Each plate contains only one type of perturbation modality.

We also considered the impact of edge effects, or plate layout effects, in our design. Edge effects are the technical artifact whereby different samples will yield different behavior depending on where they are located on a plate; generally this is observed mainly in the outer two rows and columns of the plate, and the problem persists despite efforts to mitigate it experimentally[19]. While designing the plate layout, we divided the plate into outer and inner wells, where the outer wells are the two rows and columns closest to the edge of the plate and the inner wells are the rest of the wells on the plate. Then we applied the following constraints to minimize the impact of edge effects: both of the compounds that have the same target will either be in the inner wells or in the outer wells (they will not be split such that one of the compounds is in the inner well while the other is in the outer well); the gene associated with the target of outer well compounds will be in the outer wells of the genetic perturbation plate; and all of the positive control compounds will be in the inner wells. If preferable, with this design, an analysis can be constrained to the inner wells only, to ensure that edge effects have minimal influence on the results.

Experimental conditions

Compounds. The treatment compounds were assayed at 5 µM and the cell seeding density was 1,000 cells per well.

Open reading frames. The cell seeding density was 1,625 cells per well, with a media seeding volume of 40 µl per 384 wells. The viral volume was 1 µl virus for 384 wells. The concentration of polybrene was 4 µg ml^{-1}. The media was changed after 24 h, removing polybrene and virus and adding back 50 µl media. No selection was done with blasticidin.

CRISPR. The cell seeding density was 350 cells per well, with a media seeding volume of 40 µl for 384 wells. The viral volume was 0.5 µl virus for 384 wells, and the polybrene concentration was 4 µg ml^{-1}. The media was changed after 24 h, removing polybrene and virus and adding back 50 µl media. No selection was done with blasticidin.

Experimental conditions tested

Although constrained by cost, we captured the compound, ORF and CRISPR plates under several experimental conditions to identify those that improve gene–compound, compound–compound and gene–gene matching (Fig. 3). We did more replicate plates for conditions that were less expensive or that were the most promising, and for the compound and CRISPR plates that had only singlicates of most samples (as compared with ORFs, which had duplicates on the plate). Additionally, we captured plates under many other experimental conditions, listed below, to optimize the experimental conditions. UMAP embeddings of all of the experimental conditions are shown in Extended Data Figs. 5–10.

We note that the cell types commonly used are historical lines derived from two white patients, one male (A549) and one female (U2OS). Therefore, conclusions from these data may hold true only for the demographics or genomics of those persons and not broader groups. They were chosen because the lines are both well-suited for microscopy, and they offer the advantage of enabling direct comparison with extensive prior studies using them.

Primary group of experimental conditions. For the primary group of experimental conditions we used four replicate plates of compounds and CRISPR guides and two replicate plates of ORFs (which, as mentioned, contain two replicates in each plate) at two time points and two cell lines each. The short and long time points were different for each perturbation type: for compounds they were 24 h and 48 h; for ORFs they were 48 h and 96 h; and for CRISPR guides they were 96 h and 144 h. The two cell lines were U2OS and A549. For CRISPR experiments, polyclonal A549 and U2OS were used.

Secondary group of experimental conditions. The experimental conditions are as follows: we used one A549 96 h ORF plate, in which the cells have been additionally treated with blasticidin (a drug that kills cells that have not been properly infected with the genetic reagent); two replicate plates of the A549 144 h CRISPR plate, in which the cells have been additionally treated with puromycin (a drug that kills cells that have not been properly infected with the genetic reagent); two replicate plates of the A549 48 h compound plate with 20% higher cell seeding density than the baseline; two replicate plates of the A549 48 h compound plate with 20% lower cell seeding density than the baseline; we imaged four replicate plates of the A549 24 h compound plate six additional times to test for photobleaching from repeated imaging; we imaged two replicates of the ORF plates in U2OS and A549 at 96 h and 144 h four additional times, once each on days 1, 4, 14 and 28 after the first imaging, to test the stability of samples over time; and we used four replicate plates of 48 h compound plates in polyclonal A549 cells with Cas9.

Number of plates and images

Across both of the groups of experimental conditions there are 51 physical plates and 107 plates of images: 40 physical plates and 40 plates of images in the primary group, and 11 physical plates and 67 plates of images in the secondary group. Each plate consists of 384 wells and, on average, nine sites were imaged within each well. At each site, eight (five fluorescent and three brightfield) images were captured. This amounts to nearly 3 million images across the 107 plates.

Sample preparation and image acquisition

The A549 and U2OS parent lines can be ordered from ATCC directly, but the Cas9 cells are not available due to Broad Institute licensing restrictions. The Cell Painting assay involves staining eight components of cells with six fluorescent dyes that are imaged in five channels: nucleus (Hoechst; DNA), nucleoli and cytoplasmic RNA (SYTO 14; RNA), endoplasmic reticulum (concanavalin A; ER), Golgi and plasma membrane (wheat germ agglutinin (WGA); AGP), mitochondria (MitoTracker; Mito), and the actin cytoskeleton (phalloidin; AGP) (Fig. 1). We optimized the Cell Painting assay described by ref. 6 by changing the concentrations of Hoechst and phalloidin, and combining dye addition and dye permeabilization steps to create Cell Painting v2.5. The changes to the protocol are listed at https://github.com/carpenterlab/2022_Cimini_NatureProtocols/wiki#changes-in-the-official-protocol-to-create-v25-chandrasekaran-et-al-2021. The changes are described in more detail by ref. 7, in which we further developed Cell Painting v3, where we also changed the concentrations of concanavalin A and SYTO14. The images were acquired across five fluorescent channels plus three brightfield planes using a Revvity Opera Phenix HCI microscope in widefield mode with 16 bit depth and with a ×20, 1.0 numerical aperture water immersion lens. Pixel binning was used at 2 × 2, for a final effective pixel size of 0.598 µm.

Image display

In Fig. 1, each channel was mapped to a final 0–255 look-up table per the following colors and display ranges in Fiji[20]: channel 1 (Mito): Red (1,078–10,191); channel 2 (AGP): Orange Hot (426–22,225); channel 3 (RNA): Yellow (360–33,716); channel 4 (ER): Green (238–14,272); and channel 5 (DNA): Cyan (238–20,508). The final image represents the maximum value in the RGB color space for all five channels, created with Fiji's Composite (Max) mode. No other (linear or nonlinear) adjustments were performed. Panels were assembled using the Magic Montage Fiji tool and channel annotations added in Google Draw.

Image processing

We used CellProfiler bioimage analysis software (v4.0.6) to process the images using classical algorithms[21]. We corrected for variations in background intensity[22] and then segmented cells, distinguishing between

nuclei and cytoplasm. Then, across the various channels captured, we measured various features of cells across several categories including fluorescence intensity, texture, granularity, density and location (see http://cellprofiler-manual.s3.amazonaws.com/CellProfiler-4.0.6/index.html for more details). Following the image analysis pipeline (see https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/tree/6ba3fcd1495d9e844e4607373a568641981ffcd8/pipelines for the pipelines), we obtained 5,792 feature measurements from each of more than 75 million cells.

Image-based profiling

We used cytominer and pycytominer workflows to process the single-cell features extracted using CellProfiler[23–26]. We aggregated the single-cell profiles by computing the median profile, and then normalized the averaged profiles by subtracting the median and dividing by the median absolute deviation (m.a.d.) of each feature. This was done in two ways: using the median and m.a.d. of, first, the negative control wells on the plate (used in the analysis shown here), and second, all the wells on the plate. Finally, we filtered out redundant features (such that no pair of features has Pearson correlation greater than 0.9), as well as features with near-zero variance across all the plates.

Average precision, mean average precision and fraction retrieved

Average precision serves as our primary metric for both replicability (how distinguishable the replicates of a perturbation are from negative controls) and biological relevance (how distinguishable the sister perturbations are from other perturbations). We measure the similarity between profiles using cosine similarity or absolute value of cosine similarity for cases in which both positive and negative correlations are considered matches. To calculate average precision we first construct a binary ranking of sample and negative control profiles by their cosine similarities. Based on this rank list, calculation of the average precision (AP) score follows the common formulation in the field[27], that is, we average the precision values at each rank k where recall changes:

$$\mathrm{AP} = \sum_k (R_k - R_{k-1}) P_k$$

We then assign a P value to each average precision score by a permutation-based significance testing approach. Specifically, we assess the significance of the average precision score against a null distribution built by randomly shuffling the rank list 100,000 times and computing corresponding random average precision scores. These P values are then adjusted for multiple testing using the Benjamini–Hochberg procedure to obtain corrected P values (which we refer to as q values).

We also compute mean average precision by calculating the mean of the average precision scores for each class, where 'class' refers to either a specific perturbation (for replicability) or perturbations associated with the same gene (for biological relevance). We report a per-class mean average precision value along with a combined q value obtained by taking the geometric mean of the q values. Finally, we summarize the mean average precision values for a specific task by calculating the fraction of classes with q values below the significance threshold (0.05), termed 'fraction retrieved'.

Recommended dataset splits

The methods presented in the benchmarks do not involve any training (we simply use a predetermined similarity metric and hand-engineered features or a pre-trained model) and thus did not require the typical train–validate–test data splits. Depending on the use case, we suggest using different splits. For the two benchmarks, we offer the following guidelines for creating data splits when training is involved.

Representation learning. For general representation learning and domain adaptation tasks, one could train on the dataset from one cell line or time point and test it on the other cell line or time point.

Gene–compound matching. For gene–compound matching, first, all replicates of a perturbation should be in the same split. Second, for the CRISPR dataset, both guides should be in the same split. Third, three of the compounds (BVT-948, dexamethasone and thiostrepton) have two different identifiers each in the dataset (because of small differences in structures) but the same compound name, therefore each pair should be in the same split. And last, if analyzing data at the single-cell level, all cells from a well should be in the same split.

We provide recommended data splits in https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/tree/6ba3fcd1495d9e844e4607373a568641981ffcd8/datasplits

Tools and software used

Data analysis was performed using python code written in a Jupyter notebook environment[28,29]. Python libraries used for data analysis include Numpy, Scipy, scikit-learn and pandas[30–33]. Plots were generated using matplotlib, seaborn and Plotly libraries[34–36]. The Fiji[20] Magic Montage plugin, Lucidchart and Inkscape were used for generating montages, creating schematics and for adding text to images. Two-dimensional representations of image-based profiles in Fig. 2 and Extended Data Figs. 5–10 were generated using UMAP[37].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Well-level morphological profiles, image analysis pipelines, profile generation pipelines, plate maps and plate and compound metadata, and instructions for retrieving the cell images and single-cell profiles are publicly available online at https://broad.io/cpjump1. The landing page of the GitHub repository for this dataset has relevant additional information: https://broad.io/cpjump1. Cell Painting images and single-cell profiles are available at the Cell Painting Gallery on the Registry of Open Data on AWS (https://registry.opendata.aws/cellpainting-gallery/) under accession number cpg0000-jump-pilot. For well-level aggregated profiles, we use GitHub as the hosting platform and the files are stored in GitLFS. We have released the data with a CC0 license. Source data are provided with this paper.

Code availability

The code for reproducing the benchmark results, tables and figures are available at https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/tree/6ba3fcd1495d9e844e4607373a568641981ffcd8/benchmark and https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/tree/6ba3fcd1495d9e844e4607373a568641981ffcd8/visualization. We have released the code with a BSD 3-Clause license.

References

15. Arrowsmith, C. H. et al. The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536–541 (2015).
16. Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via Cell Painting. *Elife* **6**, e24060 (2017).
17. Bray, M.-A. et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).
18. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

19. Lundholt, B. K., Scudder, K. M. & Pagliaro, L. A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* **8**, 566–570 (2003).
20. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
21. Stirling, D. R. et al. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).
22. Singh, S., Bray, M.-A., Jones, T. R. & Carpenter, A. E. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.* **256**, 231–236 (2014).
23. Methods for image-based cell profiling [R package cytominer version 0.2.2]. (2020).
24. Cimini, B. A. et al. *Broad Institute Imaging Platform Profiling Handbook*. https://github.com/cytomining/profiling-handbook
25. Way, G. *et al. Pycytominer: Data processing functions for profiling perturbations*.
26. Chandrasekaran, S. N., Weisbart, E., Way, G., Carpenter, A. & Singh, S. *Broad Institute Imaging Platform Profiling Recipe*.
27. Hirling, D. et al. Segmentation metric misinterpretations in bioimage analysis. *Nat. Methods* **21**, 213–216 (2024).
28. Loizides, F. & Schmidt, B. (eds) *Positioning and Power in Academic Publishing: Players, Agents and Agendas*: *Proceedings of the 20th International Conference on Electronic Publishing* (IOS Press, 2016).
29. Van Rossum, G. & Drake, F. L. *The Python Language Reference Manual* (Network Theory Ltd., 2011).
30. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
31. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
32. McKinney, W., Van Der Walt, S. & Millman, J. Proceedings of the 9th Python in Science Conference. Preprint at https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf (2010).
33. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
34. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
35. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**(60), 3021 (2021).
36. Plotly Technologies Inc. *Collaborative Data Science* (Plotly Technologies Inc., 2015).
37. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arxiv.1802.03426 (2018).

Acknowledgements

We thank the more than 100 scientists who have contributed to the organization and scientific direction of the JUMP Cell Painting Consortium. We acknowledge valuable input on data consistency from S. Seal and D. Figueiredo Vidal. We gratefully acknowledge a grant from the Massachusetts Life Sciences Center Bits to Bytes Capital Call program for funding the data production; and funding to support data analysis and interpretation from members of the JUMP Cell Painting Consortium, from the National Institutes of Health (NIH MIRA R35 GM122547 to A.E.C.), and from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (2020-225720 to B.A.C.). We also gratefully acknowledge the use of the Revvity Opera Phenix High-Content/High-Throughput imaging system at the Broad Institute, funded by the S10 Grant NIH OD-026839-01.

Author contributions

S.N.C., B.A.C., A.G., L.M., M.K.-A., J.G.D., B.F., A.S., M.M., D.K., J.B., D.E.R., S.E.S., S.G., S.S. and A.E.C. contributed to the experimental design. A.G., L.M., M.K.-A., B.F., M.M. and D.H. performed the laboratory experiments. S.N.C., B.A.C., N.J., A.A.K., J.A., M.H., H.S.-A. and S.S. performed data and/or image analysis. S.N.C., B.A.C., M.K.-A., S.E.S., S.G., S.S. and A.E.C. contributed to the interpretation of data. S.N.C., J.C.C., A.A.K., S.S. and A.E.C. wrote the paper and all authors contributed to the editing of the paper. B.A.C., S.S. and A.E.C. carried out administration and supervision.

Competing interests

S.S. and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C.: Recursion, SyzOnc, Quiver Bioscience; S.S.: Waypoint Bio, Dewpoint Therapeutics, Deepcell) and receive honoraria for occasional talks at pharmaceutical and biotechnology companies. D.K. and S.G. are employees of Merck Healthcare KGaA. All other authors have no competing interests.

Additional information

Extended data are available for this paper at https://doi.org/10.1038/s41592-024-02241-6.

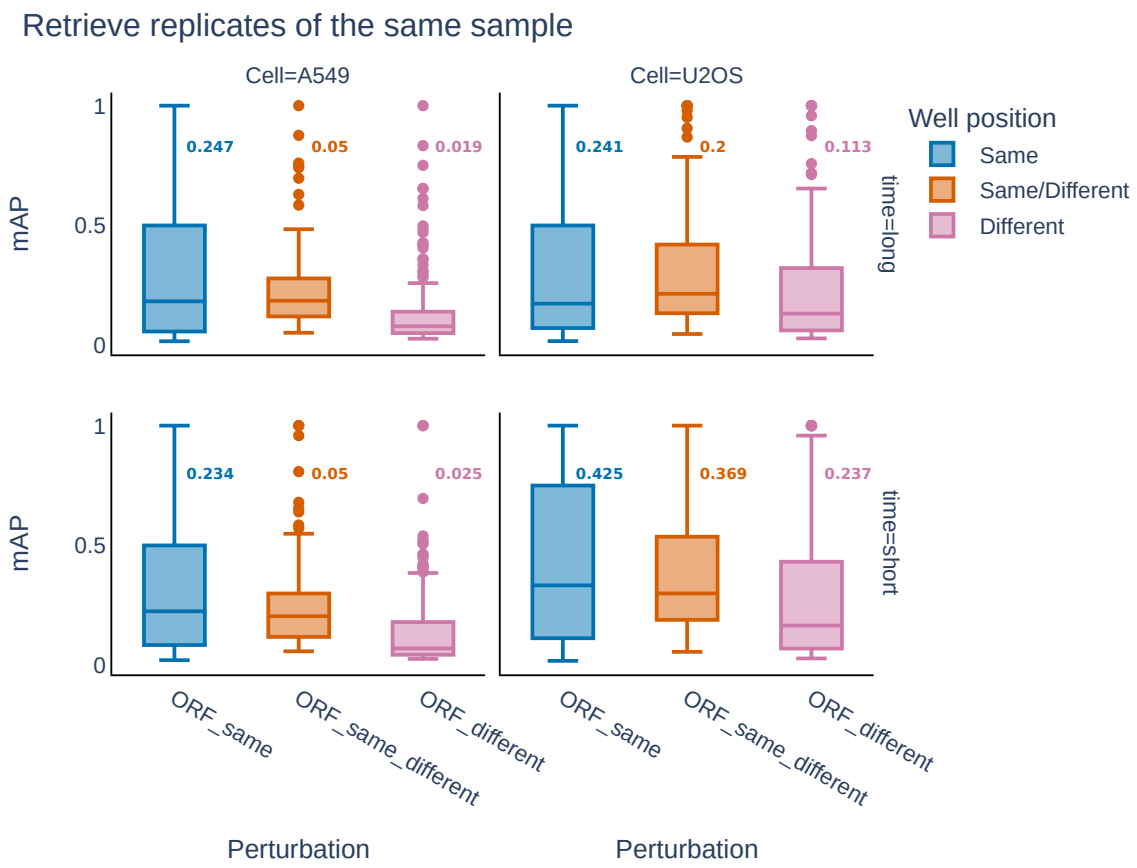
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02241-6.

Correspondence and requests for materials should be addressed to Shantanu Singh or Anne E. Carpenter.

Peer review information *Nature Methods* thanks Haiquan Li, Matthew McCall and Susanne Muller for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

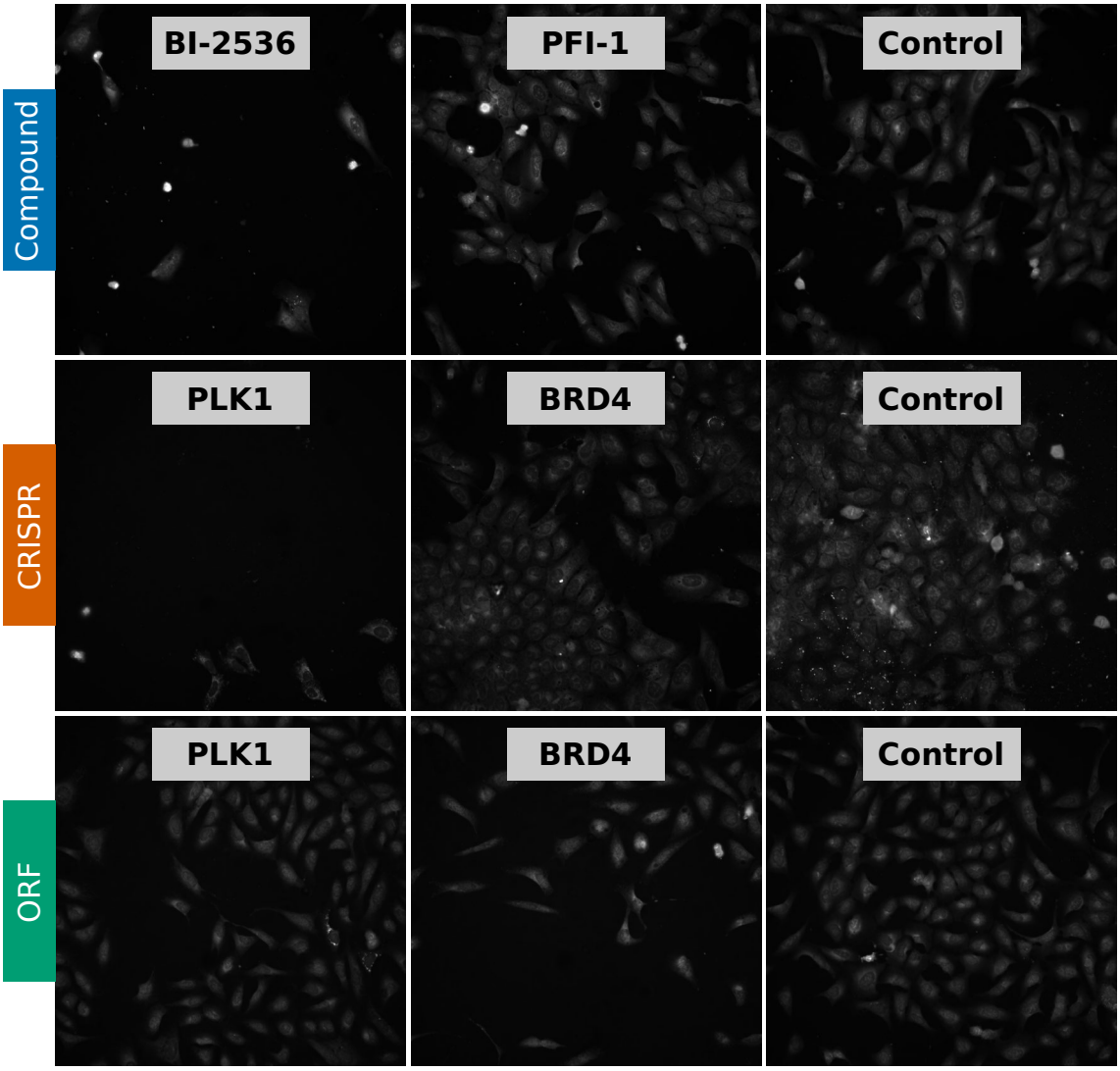
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-024-02241-6



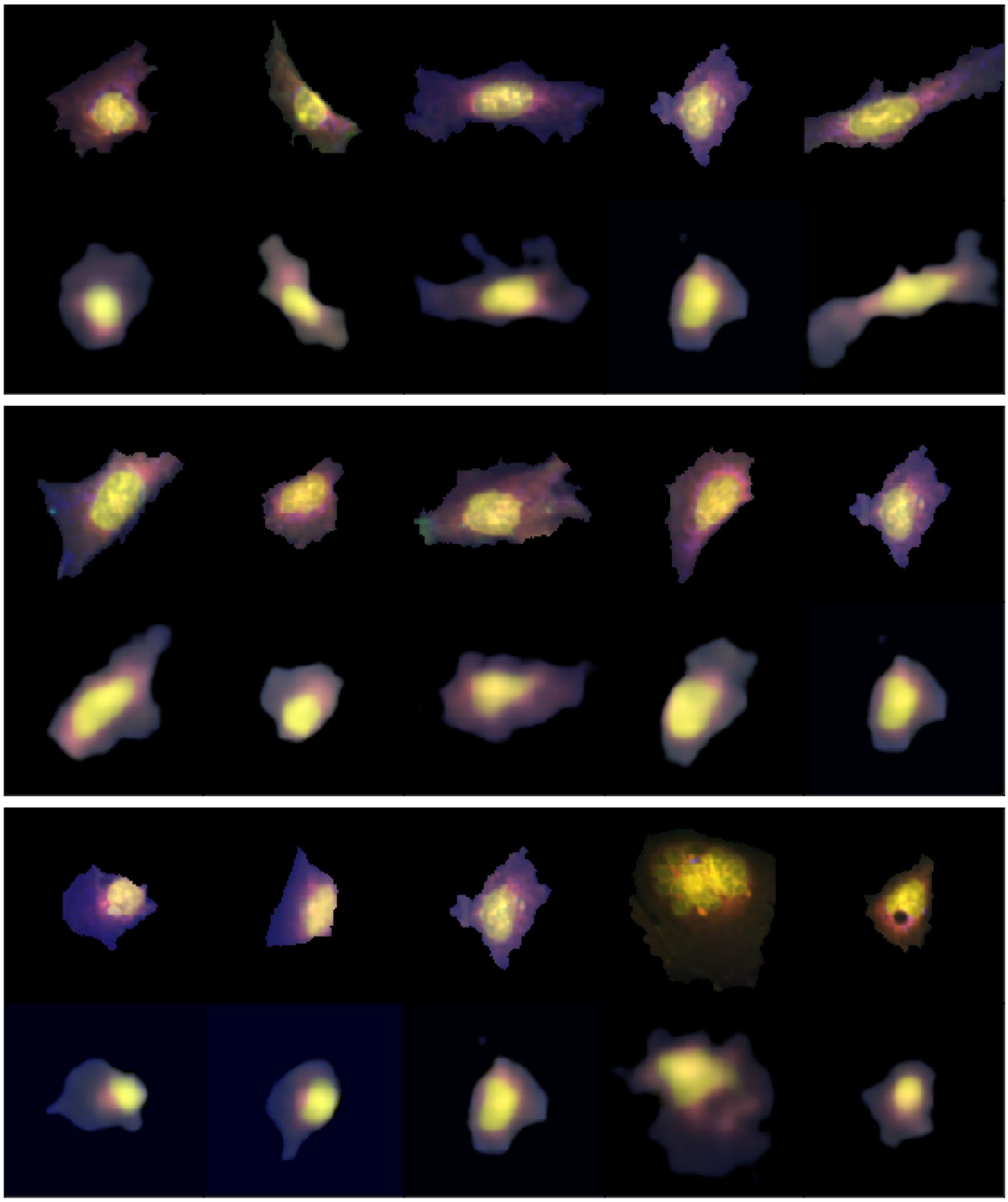
Extended Data Fig. 1 | Well position effect. mAP for perturbation detection for ORFs in the same well position (blue), same or different well positions (red) and different well positions (pink); the same or different well positions is what is shown in Fig. 4a in the main text. ORFs in different well positions are affected by plate layout effects, which lowers mAP and FR scores for retrieving replicates against a background of negative control wells. The numerical values shown above each box plot are the fraction of perturbations that can be successfully retrieved (FR) values for each retrieval task. Box plot boundaries are 75th (Q3) and 25th (Q1) percentiles, with whiskers at +/− 1.5 times the interquartile range (Q3−Q1) or the highest or lowest data point. n = 320 biologically independent ORF reagents in the blue boxes and n = 160 biologically independent ORF reagents in the pink and red boxes.

https://doi.org/10.1038/s41592-024-02241-6



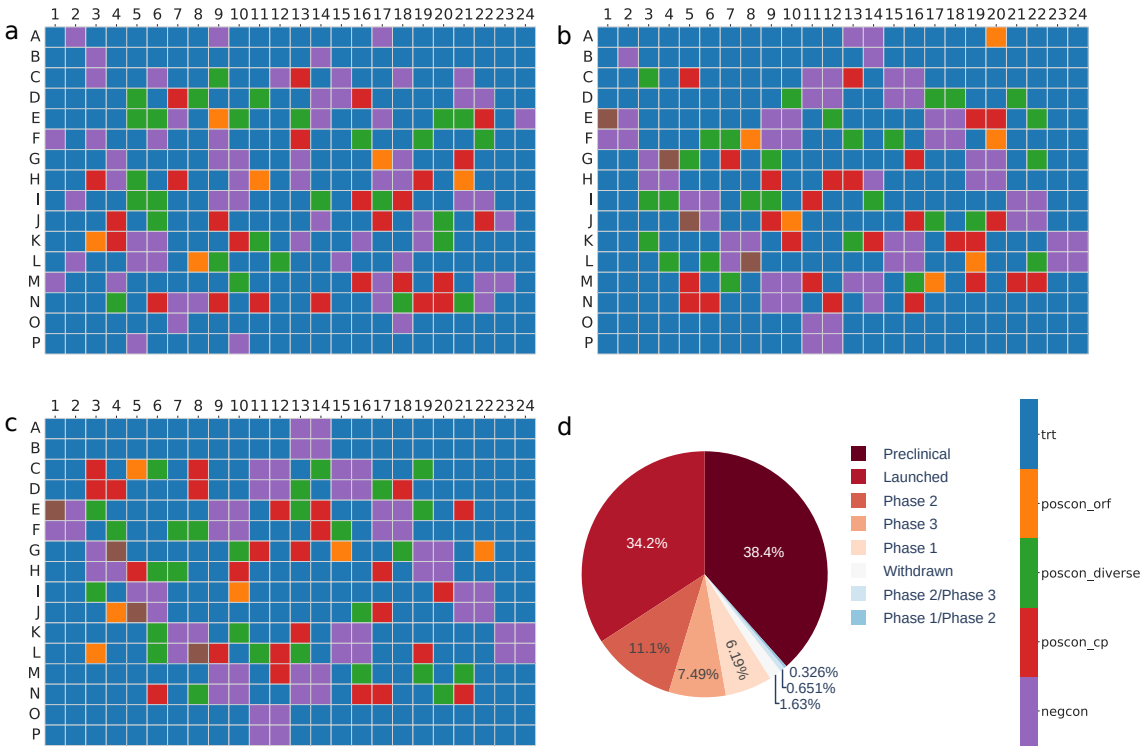
Extended Data Fig. 2 | Similarity of perturbation impact across modalities for genes and compounds related to the compound BI-2536. Treatment of U2OS cells with BRD4 inhibitors BI-2536 (multi-target, including PLK1 and BRD4) and PFI-1 (BRD4-specific) is shown (top row; all images are composites of max intensity across all five imaging channels). BI-2536 causes cell death, and this phenotype is mimicked by PLK1 knockout (middle row). In contrast, BRD4 knockout fails to produce a distinct phenotype, death-related or otherwise, as is the case for the BRD4-specific inhibitor PFI-1 (middle column, top and middle row); both profiles are quantitatively similar to negative controls. This implies that BRD4 inhibition has a limited phenotypic impact in this assay under these experimental conditions, allowing the PLK1-inhibiting phenotype of BI-2536 to dominate the profile. BRD4 overexpression, on the other hand, also induces cell death (bottom row) and a profile strongly similar to BI-2536, which could indicate that BRD4 overexpression yields a dominant negative phenotype. Overexpression of PLK1 produces a phenotype that is not distinguishable from negative control. Negative controls for compounds, CRISPR, and ORF reagents are included. These are representative images from one of the four replicate wells of each treatment in the dataset. Wells were sampled from the longer time point for each perturbation modality (Supplementary Table 1).

https://doi.org/10.1038/s41592-024-02241-6



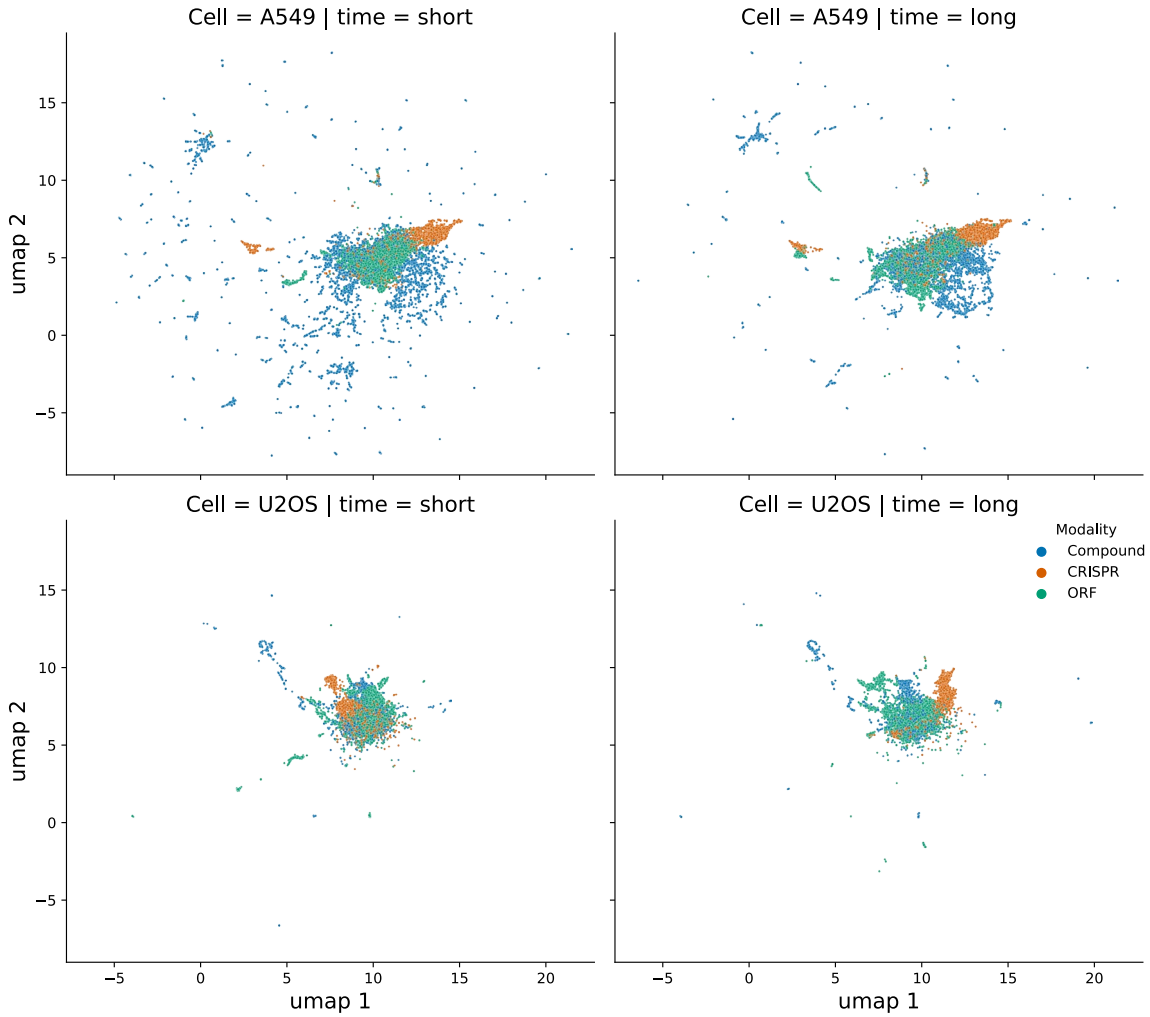
Extended Data Fig. 3 | CellProfiler features to cell images. Example single-cell images (first row) and their synthetically generated version (second row) are shown in each sub figure. The synthetic version is generated by each single cell's corresponding CellProfiler measurements. To learn a transformation function from single-cell's CellProfiler extracted features to single-cell images, 7077 single cells were randomly selected from a set of eight diverse compounds (aloxistatin, AMG900, dexamethasone, TC-S-7004, FK-866, LY2109761, NVS-PAK1-1 and quinidine) to train a convolutional neural network (CNN). The set of cell-level Cell Painting measurements was reduced to a non-redundant set of features for five channels of DNA, RNA, ER, AGP and Mito. Location related features and low variance features were excluded. Single-cell images corresponding to each cell's CellProfiler measurements were extracted by image crops of a fixed size (160 pixels) bounding box around the cell's Cells_Location_Center coordinates. CNN model learns the transformation from (3019,1) size CellProfiler features to (128, 128, 5) size images.

https://doi.org/10.1038/s41592-024-02241-6



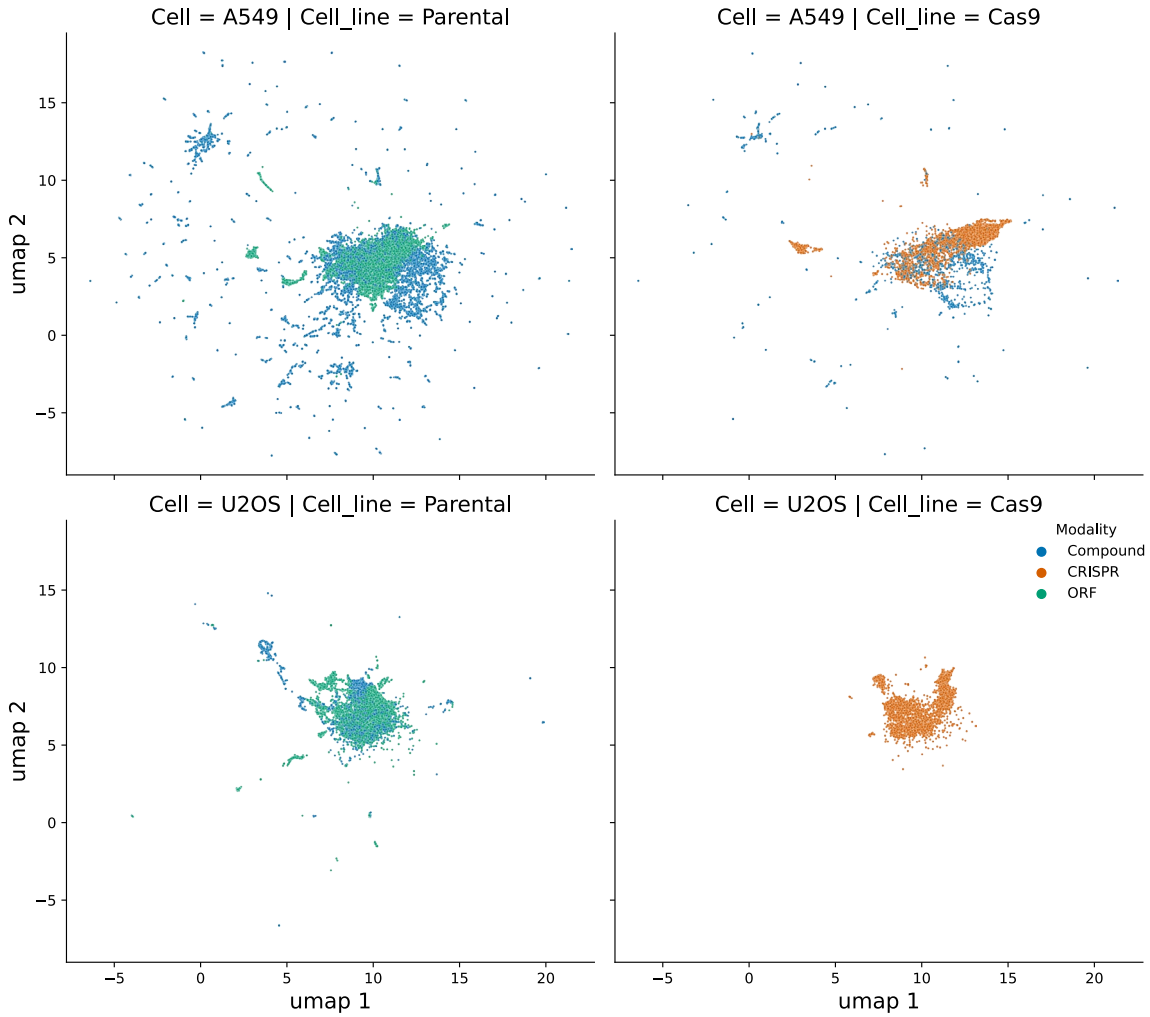
Extended Data Fig. 4 | Plate maps and overview of compounds' clinical phase status. Maps in a-c show **a**) Compound plate, **b**) CRISPR plate and **c**) ORF plate. The control wells and the treatment (trt) wells are shown in different colors. Poscon are positive controls (additional details in the Methods section) and negcon is the negative control. **d**) Over a third of the compounds in the dataset have been launched for sale, whereas others have progressed to various stages of human clinical trials.

https://doi.org/10.1038/s41592-024-02241-6



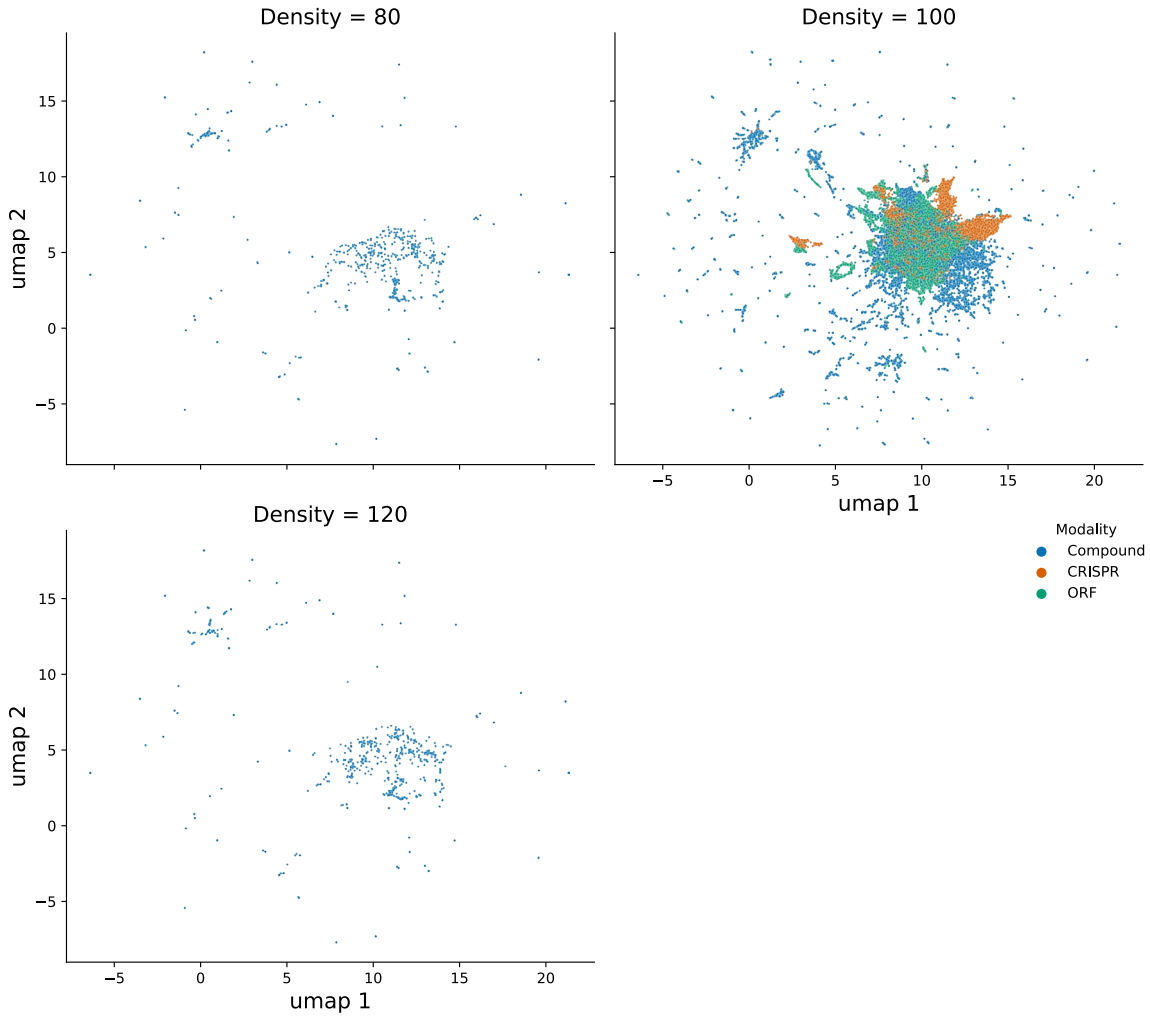
Extended Data Fig. 5 | Cell type and time, for all tested perturbations and conditions. The primary group of tested samples in the CPJUMP1 experiment consists of three perturbation modalities (compounds, CRISPR guides and ORFs), two cell types (U2OS and A549) and two time points per perturbation modality (Supplementary Table 1). This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

https://doi.org/10.1038/s41592-024-02241-6



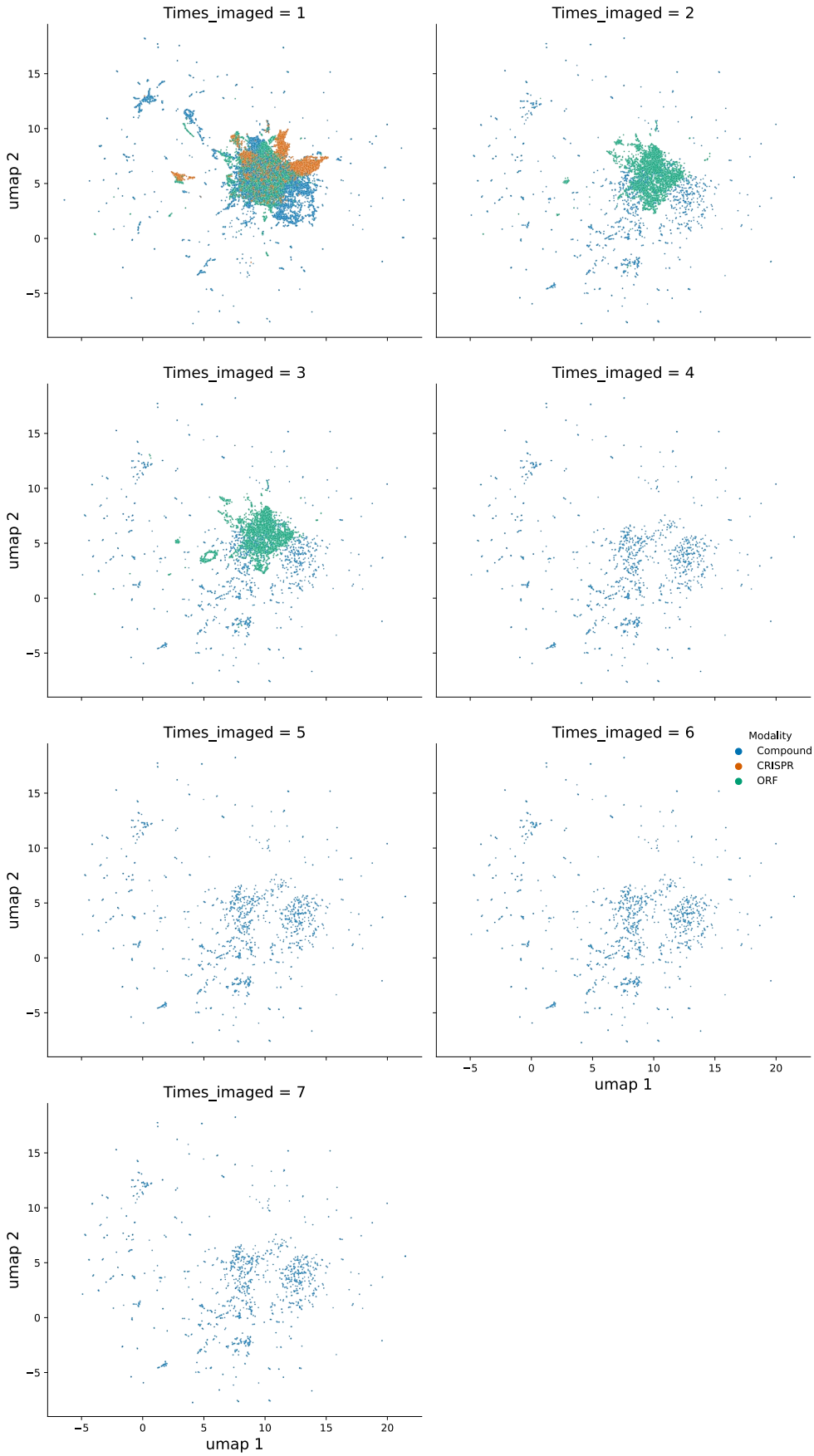
Extended Data Fig. 6 | Cas9 status. Parental line is the original cell line with no modifications. Cas9 cell line is a polyclonal cell line expressing Cas9 (used for all CRISPR and one compound experiment). This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

https://doi.org/10.1038/s41592-024-02241-6



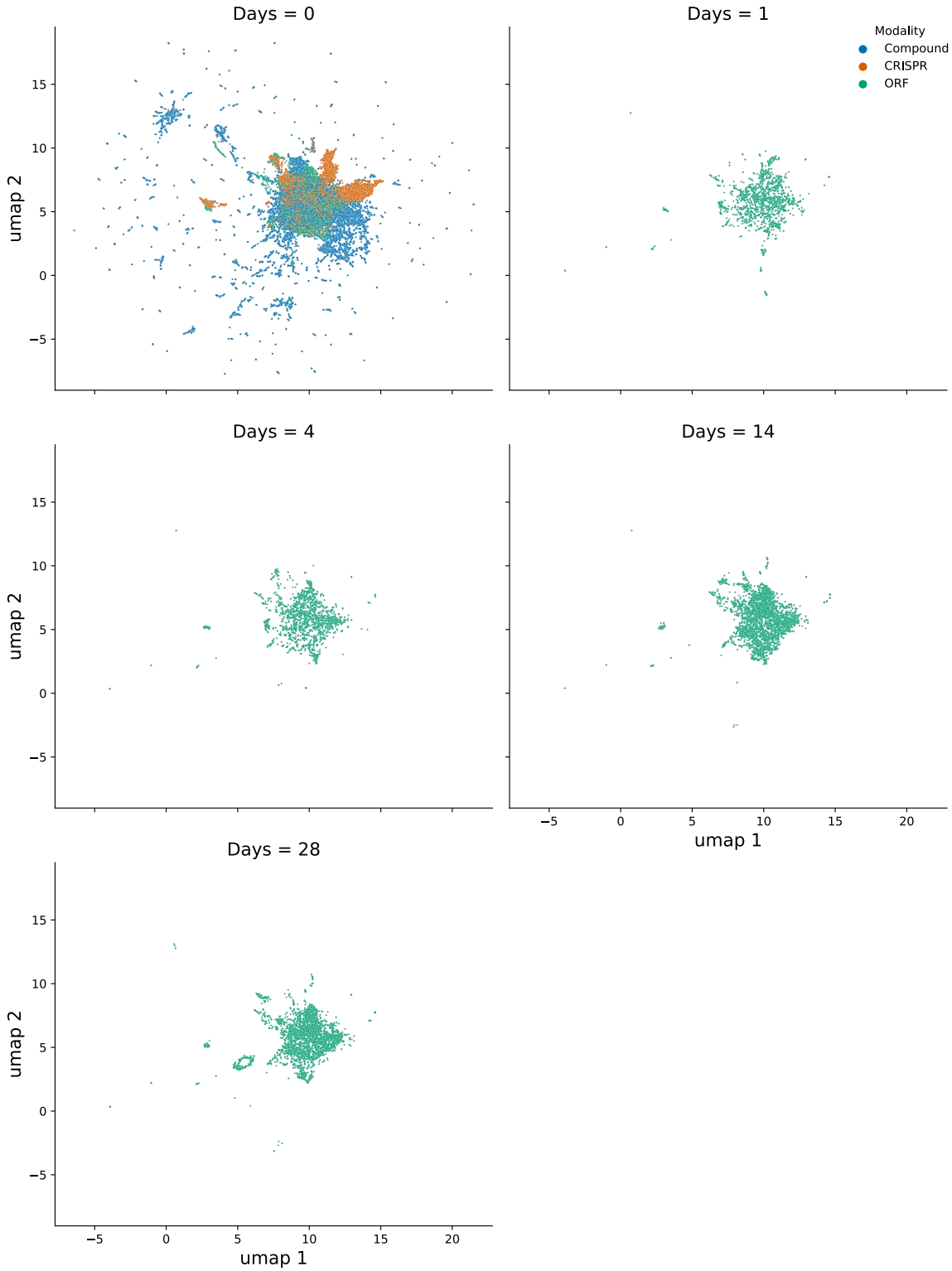
Extended Data Fig. 7 | Different cell seeding densities. Experiments were performed with the baseline seeding density (100%; 1000 cells/well), increased seeding density (120%), and decreased seeding density (80%). This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

https://doi.org/10.1038/s41592-024-02241-6



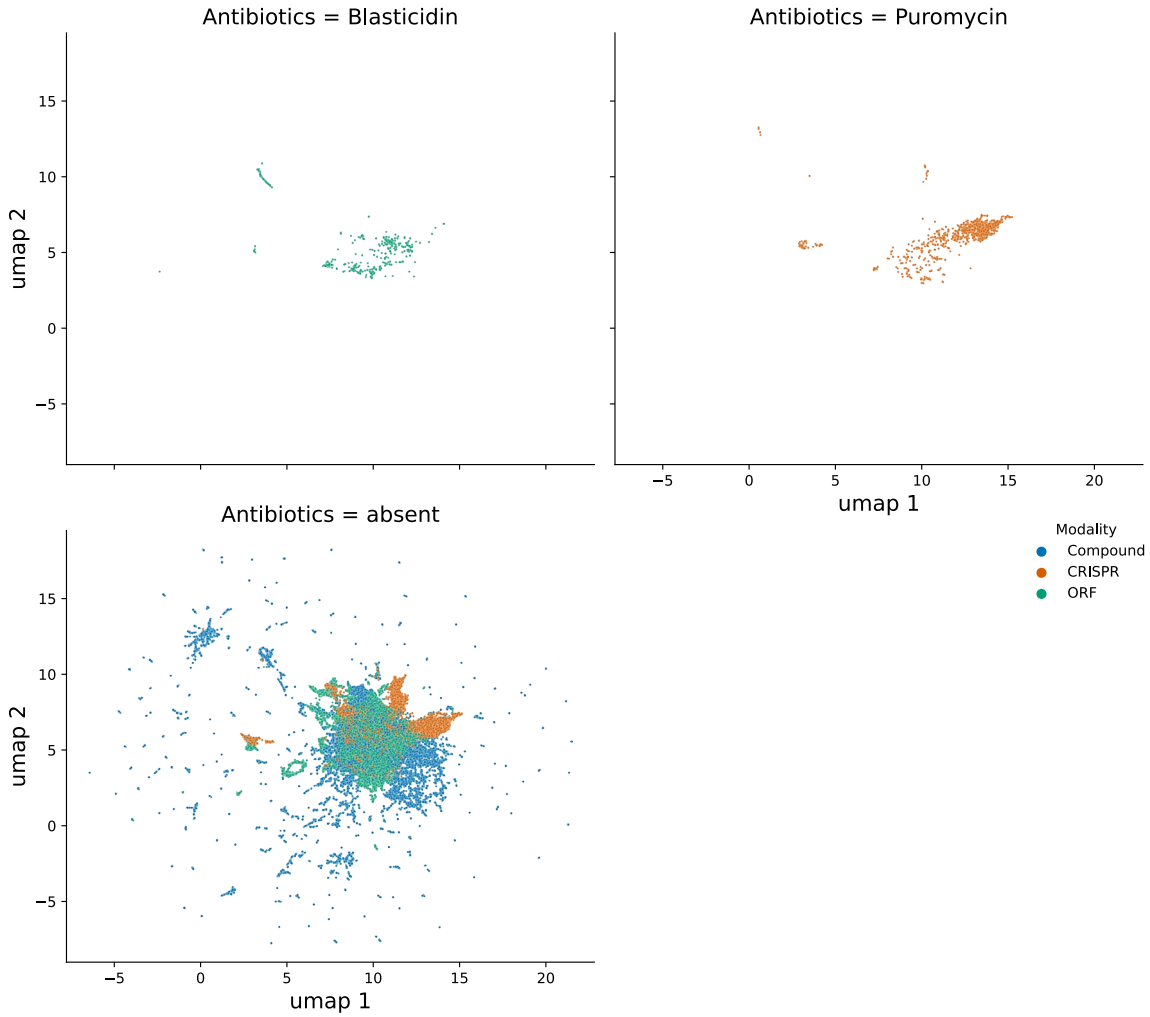
Extended Data Fig. 8 | Impact of repeat imaging. Some plates were imaged more than once. This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

https://doi.org/10.1038/s41592-024-02241-6



Extended Data Fig. 9 | Imaging after a time delay. A subset of plates were imaged after a certain number of days. This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

https://doi.org/10.1038/s41592-024-02241-6



Extended Data Fig. 10 | Antibiotic selection. In some CRISPR and ORF plates, cells were selected using antibiotics. This UMAP plot includes the CPJUMP1 primary experiment (4 Compound, 4 CRISPR and 2 ORF plates per cell type and time point) plus all other data points from the CPJUMP1 experiments, as outlined in Fig. 3.

nature portfolio

Corresponding author(s): A.E. Carpenter

Last updated by author(s): Feb 21, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	Data analysis was performed using python code written in a Jupyter notebook environment. Python libraries used for data analysis include Numpy, Scipy, scikit-learn and pandas. Plots were generated using matplotlib, seaborn and Plotly libraries. Fiji7's Magic Montage plugin, Lucidchart and inkscape were used for generating montages, creating schematics and for adding text to images. Two-dimensional representations of image-based profiles in Figs. 2 and Extended Data Figs. 5-10 were generated using UMAP . The code used in this study is available at https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods. It is available for use under the BSD 3-clause license, a permissive open-source license.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Well-level morphological profiles, image analysis pipelines, profile generation pipelines, plate maps and plate and compound metadata, and instructions for retrieving the cell images and single cell profiles are publicly available online at https://broad.io/cpjump1.

Cell Painting images and single-cell profiles are available at the Cell Painting Gallery on the Registry of Open Data on AWS (https://registry.opendata.aws/cellpainting-gallery/) under accession number cpg0000-jump-pilot. For well-level aggregated profiles, we use GitHub as the hosting platform and the files are stored in GitLFS.

We have released the data with a CC0 license.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Experiments involved cultured cells where the number of cells is governed by the size of the growth vessel; The number of replicates for various biological experiments were chosen based on each expert's knowledge of the variability in the type of experiment. We found from previous experiments that the quality of signal in the data did not improve significantly beyond four replicates.
Data exclusions	No data was excluded from the analyses.
Replication	All data presented indicates the number of replicates represented. Experiment was performed with four replicates of each treatment for each experimental condition. All replications were successful.
Randomization	We designed the experimental plates such that genes that are perturbed by CRISPR and ORF reagents, their products, are always targeted by two compounds. We randomized the location of each sample on the experimental plate as long as they satisfied the following criteria: - Both of the compounds that have the same target will either be in the inner wells or in the outer wells. They will not be split such that one of the compounds is in the inner well while the other is in the outer well. where the outer wells are the two rows and columns closest to the edge of the plate and the inner wells are the rest of the wells on the plate. - The gene associated with the target of outer well compounds will be in the outer wells of the genetic perturbation plate. - All the positive control compounds are in the inner wells.
Blinding	Experiments involved cells; group allocation is not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	A549 Catalog: CCL-185 Lot: 70018877 U2OS Catalog: HTB-96 Lot: 70016635
Authentication	Since the cells were ordered directly from ATCC, we did not authenticate via SNP or STR profiling.
Mycoplasma contamination	We tested the parental cells ordered from ATCC for mycoplasma contamination a few months into culturing them to ensure they remained negative.
Commonly misidentified lines (See ICLAC register)	To our knowledge neither of these cell lines, U-2 OS nor A549 are commonly misidentified.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2025 June 09.

Published in final edited form as:

Cell. 2024 December 12; 187(25): 7045–7063. doi:10.1016/j.cell.2024.11.015.

How to build the virtual cell with artificial intelligence: Priorities and opportunities

Charlotte Bunne[1,2,3,4,50], **Yusuf Roohani**[1,3,5,50], **Yanay Rosen**[1,3,50], **Ankit Gupta**[3,6], **Xikun Zhang**[1,3,7], **Marcel Roed**[1,3], **Theo Alexandrov**[8,9], **Mohammed AlQuraishi**[9], **Patricia Brennan**[3], **Daniel B. Burkhardt**[11], **Andrea Califano**[10,12,13], **Jonah Cool**[3], **Abby F. Dernburg**[14], **Kirsty Ewing**[3], **Emily B. Fox**[1,15,16], **Matthias Haury**[17], **Amy E. Herr**[16,18], **Eric Horvitz**[19], **Patrick D. Hsu**[5,18,20], **Viren Jain**[21], **Gregory R. Johnson**[22], **Thomas Kalil**[23], **David R. Kelley**[24], **Shana O. Kelley**[25,26], **Anna Kreshuk**[27], **Tim Mitchison**[28], **Stephani Otte**[17], **Jay Shendure**[29,30,31,32], **Nicholas J. Sofroniew**[33], **Fabian Theis**[34,35,36], **Christina V. Theodoris**[37,38], **Srigokul Upadhyayula**[14,16,39], **Marc Valer**[3], **Bo Wang**[40,41], **Eric Xing**[42,43], **Serena Yeung-Levy**[1,44], **Marinka Zitnik**[45,46,47], **Theofanis Karaletsos**[3,*], **Aviv Regev**[2,*], **Emma Lundberg**[3,6,7,48,*], **Jure Leskovec**[1,3,*], **Stephen R. Quake**[3,7,49,*]

[1]Department of Computer Science, Stanford University, Stanford, CA, USA

[2]Genentech, South San Francisco, CA, USA

[3]Chan Zuckerberg Initiative, Redwood City, CA, USA

[4]School of Computer and Communication Sciences and School of Life Sciences, EPFL, Lausanne, Switzerland

[5]Arc Institute, Palo Alto, CA, USA

[6]Department of Protein Science, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden

[7]Department of Bioengineering, Stanford University, Stanford, CA, USA

[8]Department of Pharmacology, University of California, San Diego, San Diego, CA, USA

[9]Department of Bioengineering, University of California, San Diego, San Diego, CA, USA

[10]Department of Systems Biology, Columbia University, New York, NY, USA

[11]Cellarity, Somerville, MA, USA

[12]Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY, USA

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

[*]Correspondence: tkaraletsos@chanzuckerberg.com (T.K.), regev.aviv@gene.com (A.R.), emmalu@stanford.edu (E.L.), jure@cs.stanford.edu (J.L.), steve@czbiohub.org (S.R.Q.).

DECLARATION OF INTERESTS

C.B. and A.V.R. are employees of Genentech, a member of the Roche Group. A.V.R. has equity in Roche. A.V.R. was a co-founder and equity holder of Celsius Therapeutics and is an equity holder in Immunitas. Until July 31, 2020, A.V.R. was an S.A.B. member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov. A.V.R. is a named inventor on multiple filed patents related to single-cell and spatial genomics, including for scRNA-seq, spatial transcriptomics, Perturb-Seq, compressed experiments, and PerturbView. E.L. is an advisor for the Chan-Zuckerberg Initiative Foundation, Element Biosciences, Cartography Biosciences, Pfizer, Santa Ana Bio, and Pixelgen Technologies. N.J.S. is an employee of EvolutionaryScale, PBC.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[13]Chan Zuckerberg Biohub, New York, NY, USA

[14]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA

[15]Department of Statistics, Stanford University, Stanford, CA, USA

[16]Chan Zuckerberg Biohub, San Francisco, CA, USA

[17]Chan Zuckerberg Institute for Advanced Biological Imaging, Redwood City, CA, USA

[18]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA

[19]Microsoft Research, Redmond, WA, USA

[20]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

[21]Google Research, Mountain View, CA, USA

[22]NewLimit, San Francisco, CA, USA

[23]Schmidt Futures, New York, NY, USA

[24]Calico Life Sciences LLC, San Francisco, CA, USA

[25]Chan Zuckerberg Biohub, Chicago, IL, USA

[26]Northwestern University, Evanston, IL, USA

[27]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[28]Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[29]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[30]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

[31]Seattle Hub for Synthetic Biology, Seattle, WA, USA

[32]Howard Hughes Medical Institute, Seattle, WA, USA

[33]EvolutionaryScale, PBC, New York, NY, USA

[34]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany

[35]School of Computing, Information and Technology, Technical University of Munich, Munich, Germany

[36]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany

[37]Gladstone Institute of Cardiovascular Disease, Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, USA

[38]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA

[39]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[40]Department of Computer Science, University of Toronto, Toronto, ON, Canada

[41]Vector Institute, Toronto, ON, Canada

Cell. Author manuscript; available in PMC 2025 June 09.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[42]Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, USA

[43]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

[44]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[45]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[46]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA

[47]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[48]Department of Pathology, Stanford University, Stanford, CA, USA

[49]Department of Applied Physics, Stanford University, Stanford, CA, USA

[50]These authors contributed equally

SUMMARY

Cells are essential to understanding health and disease, yet traditional models fall short of modeling and simulating their function and behavior. Advances in AI and omics offer groundbreaking opportunities to create an AI virtual cell (AIVC), a multi-scale, multi-modal large-neural-network-based model that can represent and simulate the behavior of molecules, cells, and tissues across diverse states. This Perspective provides a vision on their design and how collaborative efforts to build AIVCs will transform biological research by allowing high-fidelity simulations, accelerating discoveries, and guiding experimental studies, offering new opportunities for understanding cellular functions and fostering interdisciplinary collaborations in open science.

INTRODUCTION

The cell, the fundamental unit of life, is a wondrously intricate entity with properties and behaviors that challenge the limits of physical and computational modeling. Every cell is a dynamic and adaptive system in which complex behavior emerges from a myriad of molecular interactions. Some aspects are remarkably robust to perturbations, such as the elimination of genes or their replacement with homologs from different species. Other aspects are sensitive to even seemingly minor disruptions, such as a point mutation or an external factor that tips cells into dysfunction and disease.

To understand a cell's function, scientists have attempted to construct virtual cell models to simulate, predict, and steer cell behavior.[1–6] Building on this vision, we use the term virtual cell to define a computational model that simulates the biological functions and interactions of a cell. Existing cell models are often rule-based and combine assumptions about the underlying biological mechanisms with parameters fit from observational data. They generally rely on explicitly defined mathematical or computational approaches, such as differential equations,[7–9] stochastic simulations,[10,11] or agent-based models.[12,13] They vary in complexity and cover different defined aspects of cell biology, such as transcription and translation,[14] cytoskeletal driven cell behavior,[15,16] biochemical networks,[17] or metabolic flux.[18,19] The first whole-cell model was developed in 2012, representing all 482 genes and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

molecular functions known for an organism: the bacteria *Mycobacterium genitalium*.[8] Since this pioneering work, genome-wide models have been developed to represent other bacterial organisms, including *Escherichia coli*.[8,20–22]

Despite their widespread use in modeling biological systems, approaches to date fall short of capturing many aspects of the operations of both bacterial and more complex systems, such as human cells. Challenges include: (1) Multi-scale modeling: cells operate on multiple scales across both time and space, from atomic to molecular to cellular and histological, with functional properties emerging through nonlinear transformation from one scale to another. (2) Diverse processes with massive numbers of interacting components: cellular function encompasses numerous interacting processes, such as gene regulation, metabolic pathways, and signal transduction. Each process involves a multitude of biomolecular species, in diverse and dynamic configurations and states. (3) Nonlinear dynamics: most cellular processes are highly nonlinear, such that small changes in inputs can lead to complex changes in outputs. Thus, despite progress in modeling specific cellular processes, these factors collectively pose a substantial roadblock to the construction of a virtual cell.

Two exciting revolutions in science and technology—in AI and in omics—now enable the construction of cell models learned directly from data. These parallel revolutions provide an unprecedented opportunity for an ambitious vision of an AI virtual cell (AIVC), a multi-scale, multi-modal, large-neural-network-based model that can represent and simulate the behavior of molecules, cells and tissues across diverse states (Figure 1).

Experimentally, the exponential increase in the throughput of measurement technologies has led to the collection of large and growing reference datasets within and across different cell and tissue systems,[23–25] with data doubling every 6 months for the past several years,[26] along with the ability to couple these measurements with systematic perturbations.[27–29]

Computationally, concurrent advances in AI have enhanced our ability to learn patterns and processes directly from data without needing explicit rules or human annotation.[30,31] Such modeling paradigms have been used successfully in the biomolecular realm, for example, to predict three-dimensional (3D) molecular structures from sequences[32–34] and interactions between different molecular components.[35–38] Recent modeling methodologies in AI provide representation and inference tools that satisfy the trifecta of being predictive, generative, and queryable, which are key utilities for advancing biological research and understanding. By building on these properties, we argue that we now have the methods to develop a fully data-driven neural network-based representation of an AIVC that can accelerate research in biomedicine by enabling fast-paced *in silico* studies, as well as powerful bridges between computational methods and confirmatory wet-lab experimentation (Figure 1).

The creation of an AIVC will enable a new era of high-fidelity simulation in biology, in which cancer biologists model how specific mutations transition cells from healthy to malignant; developmental biologists forecast how developmental lineages evolve in response to perturbations in specific progenitor cells; and microbiologists predict the effects of viral infection on not just the infected cell but also its host organism. AIVCs will empower

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

experimentalists and theorists alike, by transforming the means by which hypotheses are generated and prioritized and allowing biologists to span a dramatically expanded scope, better fitting the enormous scales of biology. Although the cellular models may not always directly identify mechanistic relationships, they can be viewed as tools for effectively narrowing the search space for mechanistic hypotheses, thereby accelerating the discovery of underlying factors behind cellular function.

This perspective article is based on extensive community discussions, including a workshop hosted by the Chan Zuckerberg Initiative, and aims to ignite the formation of a collaborative research agenda for a large-scale, long-term initiative with a roadmap for developing, implementing, and deploying AIVCs. We describe a vision catalyzed by emerging advances in AI in cell biology and their application to constructing virtual representations of cells. We lay out priorities and opportunities across data generation, AI models, benchmarking, interpretation, and ensuring biological veracity and safety (Box 1). By encouraging interdisciplinary collaborations in open science—spanning academia, philanthropy, and the biopharma and AI industries—we posit that a comprehensive understanding of cellular mechanisms is within reach. AIVCs have the potential to revolutionize the scientific process, lead to the understanding of novel biological principles, and augment human intelligence to underpin future breakthroughs in programmable biology, drug discovery, and personalized medicine (Box 2).

AIVCs

Our view of an AIVC is a learned simulator of cells and cellular systems under varying conditions and changing contexts, such as differentiation states, perturbations, disease states, stochastic fluctuations, and environmental conditions (Figure 1). In this context, a virtual cell should integrate broad knowledge across cell biology. VCs must work across biological scales, over time, and across data modalities and should help reveal the programming language of cellular systems and provide an interface to use it for engineering purposes.

In particular, an AIVC needs to have capabilities that allows researchers to (1) create a universal representation (UR) of biological states across species, modalities, datasets, and contexts, including cell types, developmental stages, and external conditions; (2) predict cellular function, behavior, and dynamics, as well as uncover the underlying mechanisms; and (3) perform *in silico* experiments to generate and test new scientific hypotheses and guide data collection to efficiently expand the virtual cell's abilities.

Next, we elaborate on these key capabilities and discuss approaches for how to achieve them.

URs

An AIVC would map biological data to UR spaces (Figure 1A), facilitating insights into shared states and serving as a comprehensive reference. These URs should integrate across three physical scales—molecular, cellular, and multicellular—and accommodate contributions from any relevant modality and context (Figure 1A). This integration will allow researchers to complement new data with existing information within the AIVC,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

leveraging its extensive biological knowledge to bridge gaps between different data. Such a comparison with prior data would provide a comprehensive context for every analysis.

Importantly, the multilevel representation should generalize to new states that are not present within the data used to train the AIVC. Such an emergent capability would unlock discoveries about biological states that have not been directly observed or might not even occur in nature. For instance, the AIVC's exposure to similar instances during training, such as inflammatory states in macrophages, might enable it to predict previously unknown inflammatory states in microglia. Additionally, the AIVC should be able to predict novel states resulting from interventions (or, equivalently, interventions needed to achieve a novel specified state) offering a range of downstream applications in cell engineering and synthetic biology.

Predicting cell behavior and understanding mechanisms

A defining function of an AIVC will be its ability to model cellular responses and dynamics. By training on a wide range of snapshots, time-resolved, non-interventional, and interventional datasets collected across contexts and scales, the AIVC can develop an understanding of the molecular, cellular, and tissue dynamics that occur under natural or engineered signals. These signals include external and internal stresses or other factors such as chemical (e.g., small molecules) or genetic (engineered or natural) perturbations and their combinations. An AIVC should be able to predict responses to perturbations that have not been previously tested in the lab, while also accounting for the specific features of the cellular context within which the perturbation is being tested.

The AIVC should also have the capability to simulate the temporal evolution of alterations in cell states in response to both intrinsic and extrinsic factors, along with the resulting multicellular spatial arrangements. By modeling the transient nature of the overall cell state and the continuous flux in cellular conditions, the AIVC could uncover previously unstudied trajectories in diverse dynamic processes, such as development, maintenance of homeostasis, pathogenesis, and disease progression. Another critical challenge is understanding the molecular mechanisms underpinning observed phenotypes and trajectories. The AIVC could propose potential causal factors behind phenotypes by simulating the effects of different interventions. Through its multi-scale design, the AIVC should be able to extrapolate the basis of cellular function across scales and link intracellular processes to phenotypes at the cell and tissue level. Thus, the AIVC opens new avenues for investigating mechanisms linked to diverse phenotypes and behaviors.

Although uncovering a phenotype's causal factors may not always be feasible through computation alone, the AIVC has the potential to reduce the space of possible hypotheses. Through simulating the effects of different interventions, the AIVC could propose potential causal factors behind phenotypes with corresponding degrees of uncertainty, allowing scientists to validate claims experimentally.

In silico experimentation and guiding data generation

For real-world utility, a defining function of an AIVC will be its ability to guide data generation and experiment design. An AIVC should be queryable with computational twins

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of today's laboratory experiments, here called virtual instruments (VIs). Virtual experiments could, for example, simulate experiments in a cell type that is challenging to cultivate *in vitro* or simulate expensive readouts from low-cost measurements, such as single-cell transcriptomes from label-free imaging.[39] Virtual experiments could also be used to screen a vast number of possible perturbagens at a scale that would be impossible in the lab. Such capabilities are invaluable when considering the exponentially larger search space of combinatorial perturbations involving more than one perturbagen.[40–44]

AIVCs will usher in a new pradigm of how computational systems are probed during the design of new biological experiments. In this framework, an AIVC would not only design experiments to validate specific scientific hypotheses but also to enhance its own capabilities. Equipped with the ability to assign confidence values to its predictions, an AIVC could enable interactive querying to guide experimentalists to the most efficient path for generating additional data for fine-tuned improvement in low-confidence areas. Extended to an active and iterative lab-in-the-loop process, we envision efficient and focused expansion of the AIVC's performance. Ultimately, the AIVC might even be able to identify key gaps in its own understanding of biology and propose the most efficient paths to bridge them.[45–47]

BUILDING THE AIVC

We envision an AIVC as a comprehensive AI framework composed of several interconnected foundation models that represent dynamic biological systems at increasingly complex levels of organization—from molecules to cells, tissues, and beyond. Our approach has two main components: (1) a universal multi-modal multi-scale biological state representation and (2) a set of VIs, which are neural networks that manipulate or decode these representations. Although there may be other approaches to building an AIVC, we believe this approach would provide a scaffold that can be scaled in a collaborative and open way.

We use the term UR to refer to an embedding produced by a multi-modal AIVC foundation model. An embedding is a learned numerical representation of data in a continuous vector space. The AIVC transforms high-dimensional multi-scale multi-modal biological data into embeddings that retain meaningful relationships and patterns.

The AIVC can capture cell biology at three distinct physical scales by representing (1) molecules and their structures found within individual cells, (2) individual cells, as spatial collections of those interacting molecules and structures, and (3) how individual cells interact with one another and the non-cellular environment in a tissue. Each of these scales is represented by a distinct UR, building on abstractions generated by the previous layer, thus linking the different scales.

In the context of UR, VIs are neural networks that take URs as input and produce a desired output. We describe two types of VIs: decoder VIs (or decoders) that take a UR as an input and produce human-understandable output, for example, a cell type label or a synthetic microscope image, and manipulator VIs (or manipulators), which take a UR as an input and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

produce another UR as an output, for example, that of an altered cell state after perturbation. Because these instruments will operate over the same representations, they can be shared and reused across different use cases, experiments, and datasets. Thus, we envision that any scientist will be able to build a VI on top of a UR and share it with the community. The building of VIs that closely resemble real instruments, such as a microscope, has the potential to seed the development of instrument-specific lab-in-the-loop systems.

Building UR across physical scales

The AIVC would be a multi-scale foundation model that learns distinct representations of biological entities at each physical scale (Figure 2C). These representations can be aggregated together and transformed to produce representations at the next higher physical scale. This recurring architectural motif can be applied from the level of individual molecules to the scale of entire tissues and organs granting the model consistency across biological scales (Figure 2A). Each representation applies universally to a specific class of biological entities. This abstraction allows the virtual cell to seamlessly evolve and incorporate new data—whether from new modalities or from out-of-distribution sources—within this general framework.

In the following sections, we discuss design principles and data that could be used to construct each physical scale of the AIVC bottom-up. Although many existing machine learning architectures could be applied directly to the task of learning functional representations of cellular components (Box 3), we additionally suggest the incorporation of biological inductive biases into the design of these representations, and further modeling innovations should drive the refinement and success of these models.

Molecular scale: The first layer of the virtual cell represents individual molecular species (Figures 2A and 2C). Although there are many different classes of molecules present in a cell, a starting point for the AIVC will be to model the three types of molecules of the central dogma: DNA, RNA, and proteins. These can all be represented as sequences of characters—nucleotides or amino acids.[48–53] Such sequence data are particularly well suited for AI methods originally developed for natural language processing, such as large language models (LLMs) (Box 3). Given the high-throughput measurement capabilities for genomic sequences, there are substantial and growing amounts of training data available. This abundance of data, combined with simple objective functions (such as predicting masked letters in a sequence), provides the key ingredients for effectively training models to generate an initial molecular UR. Furthermore, a biological language model could be trained on all three modalities simultaneously, thus maximizing interoperability and training corpus size. Despite its inherent compatibility with transformers, specific considerations around masking and attention mechanisms must be addressed when applying these models to biological sequence data as opposed to natural language. Although language modeling approaches have been extensively studied for these core molecules and have proven successful for some of their chemical modifications[54] and various other molecules, such as glycans, lipids, and metabolites,[55,56] they may struggle with other molecular constituents of the cell. Such modeling difficulties might be exacerbated for data that are difficult to fit into a sequence or very small molecules. Given that the primary building blocks of these

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

entities are atoms, a neural network trained to model molecules at the atomic level[32,57] could be a more general choice for this layer. However, models with atomic resolution introduce a substantial computational burden and might be constrained by the limited availability of training data. Although atomic-based modeling is highly accurate for many static structures, it cannot yet represent the full dynamic range of chemistry that occurs at this scale. Therefore, a broader, evolutionarily informed representation such as that of sequences may be preferred.

Cellular scale: The next level of abstraction models individual cell states (Figures 2A and 2C). As cellular function is underpinned by the molecular interactions and signaling networks formed in a cell, a cellular UR can be built using representations of molecular and other (e.g., imaging) features, describing the organization and abundance of molecular components. The key step here would be to integrate learned representations of molecules with their quantities and appropriately abstracted locations and timestamps to create a unified representation of the cell.[58–60]

Data for the cellular UR consist of measurements mapped to a single-cell level, such as measurements of the transcriptome (single-cell RNA sequencing [scRNA-seq]), chromatin accessibility (scATAC-seq), chromatin modification, transcription factor binding, and proteome.[61] Imaging technologies measure cell morphology at subcellular resolution, often together with molecular information.[29,62,63] For example, fluorescence confocal microscopy can help resolve the subcellular location of the human proteome.[64] Live-cell imaging[65] enables the study of proteins in living cells using time-lapse microscopy. Cryoelectron microscopy determines biomolecular structures at near-atomic resolution.[66,67] Super-resolution microscopy offers deeper insights into molecular processes through single-molecule imaging in living systems.[68,69] Complementing imaging approaches, mass spectrometry, and proximity-dependent labeling can unveil protein-protein associations and provide deeper insights into cell structure and signaling network rewiring.[70,71]

From a model architecture perspective, vision transformers[72] or models leveraging convolutional neural networks (CNNs)[73,74] are widely applicable to biological images to model across multiple imaging channels capturing different biological features,[75,76] while being robust to distribution shift and batch variability.[77] Autoencoders and transformers have been successfully applied for learning representations for sequence-based data.[59,78,79] Using AI algorithms to integrate different data modalities collected with sequencing and imaging technologies creates a multi-view model of the cell that can be both dynamic and predictive.[80,81]

As the AIVC model grows in complexity, it is crucial to also model cellular organelles and membraneless compartments[82] as units that play specific roles within the cell. Robustly capturing the functions of these units is vital to ensure accurate predictions, mechanistic interpretability, and model generalizability.

Given their prevalence, the cellular UR will initially rely on transcriptomics measurements, whereas imaging modalities will be key for continued modeling of cellular spatial organization and dynamics.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Multicellular scale: At the third layer of abstraction, the AIVC models the organization of cells into a multicellular UR (Figures 2A and 2C). This layer allows for the exploration of how cell-cell interactions, largely governed by spatial proximity, combine into tissues, organs, and, ultimately, whole organisms. Multicellular interactions can be analyzed after tissue dissociation (such as in scRNA-seq)[83] or *in situ* in a 2D section or 3D volume, where the tissue structure is preserved. Building the AIVC will require integration across available modalities that provide spatial insights, i.e., both spatial molecular profiling, as well as non-molecular tissue imaging data.

There are multiple methods to profile the spatial location of RNA[84] and proteins[85] in cells, along with various imaging methods for select molecular species (e.g., immunohistochemistry) or with stains for tissue structure alone (e.g., hematoxylin and eosin [H&E]). Spatial molecular biology is currently a very active area of research and method development. Although publicly available data are still limited, we foresee a rapid development in this domain providing multi-omic 2D and 3D datasets. A more generalized data generation effort together with open frameworks for spatial data[86] could greatly accelerate modeling at the multicellular scale.

The relative organization of cells within a 2D tissue section and 3D tissue volume can be represented using a graph or point cloud. The multicellular UR can be derived from such data using graph-learning techniques, such as graph neural networks (GNNs) and equivariant neural networks (ENNs). For image-based data, convolutional neural networks or vision transformers can be applied (Box 3).

Predicting cell behavior and understanding mechanisms

VIs are the “tools” that operate on UR embeddings and perform various functions and tasks. By altering URs of molecules, cells, and tissues, manipulators can abstract complex dynamic processes (Figure 2B) more simply as transitions between (distributions of) their representations (Figure 2D). Similarly, decoders can take an embedding of biological entities and predict one or more concrete properties, for example, physical structure, cell type/state, fitness, expression, or drug response.

The design of a wide array of manipulators provides us with an unprecedented set of tools for modeling cell behavior and dynamics: generative AI approaches, such as diffusion models[87] or autoregressive transformers,[88] i.e., model architectures that capture heterogeneity and parameterize continuous time dynamics, can predict a future state or evolution of a cell or molecular state (Box 3).[57,89] Using integrated data from time-lapse imaging,[65] gene expression profiles,[83] and other modalities, manipulators can allow inferring the phenotypic progression from stem cell to differentiated cell, while capturing the influence of both genetic factors and environmental conditions—through learned interpolations and extrapolations between multi-scale URs of different cell states. Similarly, they allow predicting the effect of treatments on patients, given a virtual representation of a patient’s molecular profile.

Furthermore, variations in cellular URs can be linked to corresponding changes in molecular states or their spatial localization, influenced by downstream factors, such as genetic variants

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

or functional changes in proteins, which are represented in a lower scale of the AIVC. Leveraging the ability of manipulators to model temporally resolved molecular and cellular events, decoders of the AIVC could potentially identify cellular components, molecular pathways, and their interactions that contribute to each prediction and process. As such, the multi-scale design of the AIVC may unveil mechanistic hypotheses of such processes.

Despite the remarkable advancements in protein modeling, the field continues to struggle in modeling dynamic molecular processes using foundation models. There will likely be areas of cell modeling, including dynamics, which pose similar challenges. For instance, the modeling of intricate networks of transient and weak molecular interactions, which play a crucial role in rapid fine-tuning of cellular signaling and formation of cell biological features such as condensates, may pose similar challenges. Consequently, we foresee a need for advanced data collection and modeling methodologies capable of capturing the dynamics of cellular processes, akin to those encountered in protein modeling. At the same time, although some functionalities of the AIVC heavily depend on such solutions, others (e.g., certain predictive functionalities) may be successful even without them. That is one of the appealing properties of multi-modal AI models with emergent properties and why developing the AIVC now is so compelling.

In silico experimentation and guiding data generation

Manipulator VIs operating in the UR space could further enable the exploration of a broad range of hypotheses through *in silico* experiments that virtually perturb a cell model. This might be achieved by predicting changes in the URs following a perturbation prompt (Figure 2D).[40,42–44]

The design of manipulators that predict transitions in the UR upon an *in silico* input can build on conditional generative models: deep learning architectures such as conditional deep generative models[31] allow generating the desired UR based on the property or context of interest (Box 3). Here, high-throughput perturbation screens—based on RNA-seq,[28,83,90] optical pooled screens (OPS),[29,39,91] or other technologies—offer a rich resource through which the AIVC can be trained to predict these effects. By conditioning on specific perturbations—such as environmental changes, genetic mutations, or chemical treatments—the generative model might produce a new UR reflecting the predicted cellular response. This conditioning could be achieved through learned or pre-computed embeddings of the affected molecular targets. Chemical compounds, small molecules, and metabolites could be embedded based on their chemical properties. Additionally, LLMs trained on comprehensive scientific literature and biological databases, such as Gene Ontology or drug banks, could further provide a rich contextual background used for conditioning the generative model, e.g., through considering wide range of interactions and side effects.

VIs can be designed so that predictions are accompanied by estimates of model uncertainty.[92] Under a Bayesian formulation of its predictive function, the predictions made for cell perturbation outcomes could include an uncertainty score, either implicitly via inference, deep kernels,[93,94] or through explicit estimation of the full posterior over model parameters.[95,96] Some practical approaches utilize model ensembles[97] or conformal predictions.[98,99] By assigning specific confidence levels to its predictions, the AIVC can call

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

methods for computing the expected value of additional data or approximations referred to in machine learning as active learning to guide experimental data collection[45] for expanding its UR. Alternatively, methods for computing the expected value of information could be used to guide data generation with the goal of optimizing a desired biological property.[92] Lastly, through its ability to conduct *in silico* experiments and suggest additional informative experiments, the AIVC could become an integrative part of lab-in-the-loop schemes. This allows not only for a seamless experimental validation of its predictions but also a sequence of experiments, predictions, and generations of hypotheses that gradually improve our systematic understanding of molecular circuits that drive biological functions.

DATA NEEDS AND REQUIREMENTS

A key consideration for the AIVC is which datasets and modalities must be collected to enable its effective construction. Unlike traditional experimental design, where data are generated to test specific scientific hypotheses, data collection for training the AIVC should be focused on ensuring the broad applicability and generalizability expected of the AIVC. To meet these ambitions, data would ideally span different domains and modalities, capture the heterogeneity and diversity of biological variability, and enable models to distinguish between technical (measurement) noise, stochastic biological variation, and physiological differences.

Data generation will require simultaneous exploration of temporal and physical scales, while allowing for system perturbations. Here, classical imaging technologies,[65,100,101] including live-cell, and newer structural imaging technologies, such as cryoelectron tomography and soft X-ray tomography,[66,102,103] as well as novel spatial omics technologies,[104,105] offer opportunities to model biomolecules and functions across scales. Furthermore, biological processes span a vast range of timescales, from the fastest reactions happening in picoseconds to a cell division progressing over hours to a day, tumor development occurring over years, and neurodegeneration over decades. The recent construction of universal cell atlases[101,106] may serve as a powerful resource for modeling cellular behavior over longer timescales, such as tissue formation. New approaches will be needed to build comparable datasets that capture the behavior of cells on shorter timescales, e.g., through methods such as live-cell imaging. Besides molecular measurements, an important aspect of data collection will lie in the measurement of biophysical and biochemical cellular properties to provide boundaries of physical and chemical realism to the AIVC.

Another important driver for the development of AIVCs will be multi-modal datasets. For example, datasets that bridge molecular and spatial scales, such as single-cell transcriptomics data combined with histology to understand how cells interact and what molecular signatures underpin the formation of specialized spatial niches.[107] Further technological development is needed to collect multi-modal data that better capture the relationship between molecular signatures, cell behavior, cellular regulation, and organization.

Although a core interest of virtual cell modeling will focus on human datasets for the purpose of understanding disease and aiding the development of novel therapeutics, human

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

datasets are limited in our ability to perform controlled experimentation and perturbations *in vivo*.

Here, the field of 3D tissue biology, including culture systems, such as organoids, is emerging as a tool to study the complexities of tissue architecture and function[108] in a 3D environment, while allowing perturbations of the system. Another critical avenue to surpass this limitation will be to perform diverse, organism-wide profiles of species spanning evolutionary history, across perturbations and under various conditions.[109–111] Ideally, large datasets could be collected across all three physical scales, allowing the AIVC to extend beyond disease research into other areas such as industrial biotechnology, agricultural biotechnology, infectious diseases, and climate change. However, based on data collection trends for the cellular and multicellular scales, modeling animal cells remains the most realistic.

Finally, a key aspect of biological data generation will be the exploration of combinatorial spaces: biological spaces are commonly high dimensional, and enumerating their variants is intractable in general, e.g., when considering all possible variants of a genome. Even for combinations of a small number of entities, exemplified in the case of enumerating pairs or sets of perturbations,[47,90] experimental design becomes exceedingly challenging. Because combinatorial possibilities quickly expand well beyond what is practical experimentally, or even computationally, new methods for their exploration must be developed.

How much data are needed to build the AIVC?

The scale of raw biological data is undeniable, but so is the sheer nominal size of even one human cell system, making first principle estimates challenging. For instance, the Short Read Archive of biological sequence data holds over 14 petabytes of information,[112] which is more than 1,000 times larger than the dataset used to train ChatGPT.[113] Large parts of these data may be redundant or have diminishing returns if used for training, and the scaling laws for models' performances must be investigated thoroughly.

In addition to data size, data diversity and quality are critical to ensure model performance.[114] Data from humans and model organisms, such as mice and *Escherichia coli*, are unequally represented in sequence and literature databases, which when used for training, encode strong species biases.[114] Other biases, for example, in terms of sex, specific diseases, or human ancestral populations could also reduce the impact of AIVC models.[115]

Although efforts on the data side are required, the AI models driving the AIVC must be designed to withstand and adapt to these challenges, i.e., exhibit robustness in their ability to integrate datasets of various origins and quality. This is crucial given both the rapid pace of advances in lab technologies (which preclude standardization on a single platform) and the broad diversity of modalities and cell systems that must be encompassed by the AIVC. As virtual cell efforts mature, the dialog between the scientists who develop models, those who generate experimental data, and funding organizations must be further intensified.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MODEL EVALUATION

A more important question for the development of AIVCs may not be “how do we build them?” but rather “how do we build trust in their competence and fidelity?” To this end, a comprehensive and adaptable benchmarking framework will be needed. Although various frameworks already exist for tackling specific biological questions (for example, protein structure prediction models[89] were developed in the context of the CASP evaluation framework), the AIVC will need to demonstrate generalizability across numerous biological contexts and downstream tasks. It must account for dynamic distributions that evolve due to environmental changes, infections, genetic variants, and other such factors causing distribution shifts.[116]

Even beyond generalizability, emergent capabilities, such as those associated with LLMs, could enable AIVC models to extrapolate to truly out-of-distribution data. In a biological context it may be difficult to decide how this boundary is defined during evaluation. New molecules, new cell states, and even new species could be considered within the training distribution. A new molecule could have homologs, including remote homologs, within the dataset. A new cell type or state could execute gene programs and regulatory networks found in existing cell types. A new strain could be closely related to existing species in the training data or live in similar environmental niches. Extrapolation to new data could then be limited to consider only the design of biological entities that do not naturally occur. This type of evaluation is already considered within the molecular design space because language-model-created proteins, such as esmGFP[52] or OpenCrispr1,[53] highlight how different they are from any of their naturally occurring counterparts. If extrapolation is a goal when designing these models, it is possible that additional inductive biases, fine tuning, or preference optimization using biomechanical, physics-based, or mechanistic modeling[117] would prove necessary.

The evaluation of AIVCs should prioritize both generalizability, as well as discovering new biology. Generalizability measures how well the model performs in unseen contexts, such as novel cell types and genetic backgrounds. It can be evaluated through a cross-modal reconstruction task, such as predicting gene expression given the morphology of a previously unseen cell or the next image in a sequence of microscopy images of cell state. Assessing generalizability builds confidence in the AIVC’s ability to capture core biological processes and understand how they vary across different contexts. Establishing such cross-modal benchmarks to link scales and modalities in cell biology is of imminent priority to the research community because these tasks are both biologically useful and well defined.

Ultimately, AIVC models should be judged on their ability to unlock new ways of understanding biology. Such an evaluation will ensure that model development is aligned with biological relevance. The most useful initial accomplishments will likely be to generate valuable testable hypotheses. For this purpose, validation datasets that are related to phenotypes that are experimentally verifiable may be suitable, such as growth rate of cells, molecular profiles, disrupted protein-protein interactions, or transcription factor binding.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

As the capabilities of AIVCs improve, we must consider whether statistical measures of performance are adequate, or if interpretability and biological causality would be core requirements.

INTERPRETABILITY AND INTERACTION

One of the hallmarks of scientific discovery in biology has been the creation of mechanistic models of a phenomenon under observation. When creating virtual cells, we may have to forgo our ability to build fully mechanistic models in favor of learning interactions that will generalize from data and predict beyond the observations. However, it is still desirable to strive toward increased interpretability.

Every AIVC prediction could be substantiated with the corresponding multi-scale interactions that determine resulting states, e.g., understanding how a cellular subsystem or protein complex is disrupted in a diseased tissue can aid development of therapeutic interventions.[118,119] The modular structure of the AIVC will enable researchers to pinpoint specific genes, proteins, or molecular processes involved in each predicted behavior. Patterns in the wiring of large models can also be leveraged to uncover combinatorial biological interactions, such as those between proteins, which can be projected to interpretable spaces without restricting the generality of the original model. Although many capabilities of the AIVC rely on predictive tasks, generating mechanistic hypotheses could provide experimental routes to understand and explore the AIVC's predictions further and will be vital for the adoption and use of AIVCs.

Ultimately, it will be of key interest to build an interactive layer for the AIVC that enables researchers of varying expertise to grasp and utilize its predictions effectively. AI agents, built using LLMs, could serve as virtual research assistants, providing an intuitive interface for non-experts.[46,120] Leveraging their extensive knowledge of scientific literature, these language models can offer deeper insights into the predictions made by the AIVC.

AN OPEN COLLABORATIVE APPROACH

Creating an AIVC requires tremendous investment, diverse backgrounds, and many iterations and can only be advanced by a concerted open science effort. As a scientific community, we must strive to ensure that both the development and usage of virtual cells are accessible and responsive to the entire scientific community. These efforts would greatly benefit from open data resources and data standards, a collaborative platform for cell modeling, and, especially, open benchmark datasets and common validation strategies to ensure their biological fidelity and real-world utility. Such a collaborative program could greatly accelerate progress across individual efforts and unify scientific research at a global scale, connecting myriad smaller-scale efforts.

To achieve this, multiple key parameters need to be considered. First, we must ensure that AIVCs represent and benefit all of humanity, with open data that captures human ancestral, sex, and geographic diversity.[121] Ensuring that such datasets reflect human diversity, while safeguarding individuals' privacy is a principal challenge. Second, as the size of AIVC models increases, the cost of training, fine tuning, or using them as is will also grow.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Investments in diverse data collection, infrastructure, and a platform for hosting virtual cell models will be critical to ensure representation, accessibility, and benefit to the broader scientific community. The platform should foster open and collaborative development of AIVCs, enabling active collaboration between biologists, clinicians, statisticians, and computer scientists. This platform should facilitate swift iterations between the lab and the modeling environment and offer opportunities to quickly test and benchmark new models. Third, synergistic collaboration among stakeholders is needed across the biomedical ecosystem, including philanthropy, academia, biopharma, and the AI industry. Pre-competitive collaborations can greatly accelerate our collective progress toward creating AIVCs. Besides the synchronization with data generators and other modeling efforts, collaboration with regulatory authorities and bioethics experts are crucial for benchmarking and establishing new norms that will expedite the deployment of AIVCs, while complying with legal requirements and setting standards for ethical issues for responsible use of virtual cells.

This article is intended to serve as a primer for the formation of a collaborative research agenda and roadmap for a large-scale, long-term initiative for developing and implementing AI-powered VCs. If successful, such interactive AIVC models, capable of simulating cellular biology, have the potential to fundamentally change how cell biology research is done. We foresee a future where AIVC platforms function as open, interconnected hubs for collaborative development and broad deployment of cell models to researchers but also as education hubs delivering training to researchers, as well as providing engagement activities for educators, patients, and the public.

OUTLOOK AND REASONS FOR OPTIMISM

The genetics and genomics communities have created large reference datasets, such as the human genome project,[23] HapMap,[122] the Cancer Genome Atlas (TCGA),[123] ENCODE,[124] the Genotype-Tissue Expression (GTEx) project,[125] the Human Protein Atlas (HPA),[64,126] the Human Cell Atlas (HCA),[24] and a growing number of deeply phenotyped, population-scale biobank efforts.[127] Thanks to these projects, massive reference data are now available to train machine learning models. Although these efforts will continue to grow, they also catalyze a new, parallel effort: creating a virtual simulation of cell biology, a new process for scientific inquiry.

The result, the AIVC has the potential to revolutionize the scientific process, leading to future breakthroughs in biomedical research, personalized medicine, drug discovery, cell engineering, and programmable biology. Acting as a virtual laboratory, the AIVC could facilitate a seamless interface between data derived from *in silico* experimentation and results from physical laboratories. As such, we expect the AIVC to contribute to a more unified view of biological processes, fostering alignment among scientists on how emergent properties in biology arise.

By bridging the worlds of computer systems, modern generative AI and AI agents, and biology, the AIVC could ultimately enable scientists to understand cells as information processing systems and build virtual depictions of life. As the AIVC expands the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

understanding of cellular and molecular systems, it will also increasingly allow us to program them and design novel synthetic ones. AI models have already been used to design new CRISPR enzymes,[53] functional proteins,[128] and even entire prokaryotic genomes.[51] The rapid progress in the precision of cell and genome engineering tools will accelerate this shift and different instantiations of the AIVC will compete in their ability to engineer new, functional biology capabilities as much as in their ability to represent and simulate biology.

Finally, we staunchly advocate the role for open science approaches, where the scientific community readily shares data, models, and benchmarks, where findings and insights are contextualized, and where a climate of perpetual improvement is fostered. We welcome and encourage all stakeholders across sectors and domains to engage in this endeavor. With a massive scientific undertaking and shared goals, open sharing of insights, and the power of safe, ethical, and reliable AI, we believe that we are stepping into a new era of scientific exploration and understanding. The confluence of AI and biology, as encapsulated by AIVCs, signals a paradigm shift in biology and shines as a beacon of optimism for unraveling multiple mysteries of the cell.

ACKNOWLEDGMENTS

We thank Rok Sosič, Charilaos Kanatsoulis, Lata Nair, Kexin Huang, Hanchen Wang, Minkai Xu, Michael Bereket, Romain Lopez, Takamasa Kudo, Ayush Agrawal, Arnuv Tandon, Mika Jain, Michihiro Yasunaga, Tim Jing, Michael Moor, George Crowley, Maria Brbić, Andrew Tolopko, Ivana Jelic, Ana-Maria Istrate, Sara Simmonds, Maximilian Lombardo, Pablo Garcia-Nieto, Mike Lin, Noorsher Ahmed, William Leineweber, Jan N. Hansen, Orit Rozenblatt-Rosen, Gita Mahmoudabadi, Zoe Piran, Adam Gayoso, Wei Ouyang, and Anshul Kundaje for discussions. J.L. was supported by NSF under nos. OAC-1835598 (CINES), CCF-1918940 (Expeditions), and DMS-2327709 (IHBEM); Stanford Data Applications Initiative; Wu Tsai Neurosciences Institute; Stanford Institute for Human-Centered AI; Chan Zuckerberg Initiative; Amazon; Genentech; GSK; Hi- Tachi; SAP; and UCB. E.L. was supported by Schmidt Futures, the Bridge2AI Program (NIH Common Fund; OT2 OD032742), the Cancer Cell Map Initiative (NCI Center for Cancer Systems Biology; U54 CA274502), the Wallenberg Foundation (2021.0346), Stanford Institute for Human-Centered AI, Chan Zuckerberg Initiative, Phil & Penny Knight Initiative for Brain Resilience at the Wu Tsai Neurosciences Institute, Danaher, and Param Hansa Philanthropies.

REFERENCES

1. Slepchenko BM, Schaff JC, Macara I, and Loew LM (2003). Quantitative cell biology with the Virtual Cell. Trends Cell Biol. 13, 570–576. [PubMed: 14573350]
2. Johnson GT, Agmon E, Akamatsu M, Lundberg E, Lyons B, Ouyang W, Quintero-Carmona OA, Riel-Mehan M, Rafelski S, and Horwitz R (2023). Building the next generation of virtual cells to understand cellular biology. Biophys. J 122, 3560–3569. [PubMed: 37050874]
3. Marx V (2023). How to build a virtual embryo. Nat. Methods 20, 1838–1843. [PubMed: 38057522]
4. Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, and Karr JR (2018). Emerging whole-cell modeling principles and methods. Curr. Opin. Biotechnol 51, 97–102. [PubMed: 29275251]
5. Georgouli K, Yeom J-S, Blake RC, and Navid A (2023). Multi-scale models of whole cells: progress and challenges. Front. Cell Dev. Biol 11, 1260507. [PubMed: 38020904]
6. Marucci L, Barberis M, Karr J, Ray O, Race PR, de Souza Andrade M, Grierson C, Hoffmann SA, Landon S, Rech E, et al. (2020). Computer-aided whole-cell design: Taking a holistic approach by integrating synthetic with systems biology. Front. Bioeng. Biotechnol 8, 942. [PubMed: 32850764]
7. Lauffenburger DA, and Linderman JJ (1996). Receptors: models for binding, trafficking, and signaling (Oxford University Press).
8. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, and Covert MW (2012). A whole-cell computational model predicts phenotype from genotype. Cell 150, 389–401. [PubMed: 22817898]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

9. Mangan S, and Alon U (2003). Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100, 11980–11985. [PubMed: 14530388]
10. Zopf CJ, Quinn K, Zeidman J, and Maheshri N (2013). Cell-cycle dependence of transcription dominates noise in gene expression. PLoS Comput. Biol 9, e1003161. [PubMed: 23935476]
11. Eling N, Morgan MD, and Marioni JC (2019). Challenges in measuring and understanding biological noise. Nat. Rev. Genet 20, 536–548. [PubMed: 31114032]
12. Hellweger FL, Clegg RJ, Clark JR, Plugge CM, and Kreft J-U (2016). Advancing microbial sciences by individual-based modelling. Nat. Rev. Microbiol 14, 461–471. [PubMed: 27265769]
13. Gorochowski TE (2016). Agent-based modelling in synthetic biology. Essays Biochem. 60, 325–336. [PubMed: 27903820]
14. Thiele I, Jamshidi N, Fleming RM, and Palsson B (2009). Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput. Biol 5, e1000312. [PubMed: 19282977]
15. Odell GM, and Foe VE (2008). An agent-based model contrasts opposite effects of dynamic and stable microtubules on cleavage furrow positioning. J. Cell Biol 183, 471–483. [PubMed: 18955556]
16. Popov K, Komianos J, and Papoian GA (2016). MEDYAN: mechanochemical simulations of contraction and polarity alignment in actomyosin networks. PLoS Comput. Biol 12, e1004877. [PubMed: 27120189]
17. Burke PEP, Campos CBL, Costa LDF, and Quiles MG (2020). M. G. A biochemical network modeling of a whole-cell. Sci. Rep 10, 13303. [PubMed: 32764598]
18. Li G, Liu L, Du W, and Cao H (2023). Local flux coordination and global gene expression regulation in metabolic modeling. Nat. Commun 14, 5700. [PubMed: 37709734]
19. Fang X, Lloyd CJ, and Palsson BO (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. Nat. Rev. Microbiol 18, 731–743. [PubMed: 32958892]
20. Stevens JA, Grünewald F, van Tilburg PAM, König M, Gilbert BR, Brier TA, Thornburg ZR, Luthey-Schulten Z, and Marrink SJ (2023). Molecular dynamics simulation of an entire cell. Front. Chem 11, 1106495. [PubMed: 36742032]
21. Maritan M, Autin L, Karr J, Covert MW, Olson AJ, and Goodsell DS (2022). Building structural models of a whole mycoplasma cell. J. Mol. Biol 434, 167351. [PubMed: 34774566]
22. Ahn-Horst TA, Mille LS, Sun G, Morrison JH, and Covert MW (2022). An expanded whole-cell model of E. coli links cellular physiology with mechanisms of growth rate control. npj Syst. Biol. Appl 8, 30.
23. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. (2001). The sequence of the human genome. Science 291, 1304–1351. [PubMed: 11181995]
24. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. (2017). The human cell atlas. eLife 6, e27041. [PubMed: 29206104]
25. CZI Single-Cell Biology Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Batuhan C, Jim C, Chambers S, et al. (2023). CZ CELL×GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at bioRxiv.
26. Heimberg G, Kuo T, DePianto D, Heigl T, Nathaniel D, Salem O, Scalia G, Biancalani T, Turley S, Rock J, et al. (2023). Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. Preprint at bioRxiv.
27. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 167, 1853–1866.e17. [PubMed: 27984732]
28. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, Pliner HA, Jackson DL, Daza RM, Christiansen L, et al. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. Science 367, 45–51. [PubMed: 31806696]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

29. Feldman D, Funk L, Le A, Carlson RJ, Leiken MD, Tsai F, Soong B, Singh A, and Blainey PC (2022). Pooled genetic perturbation screens with image-based phenotypes. Nat. Protoc 17, 476–512. [PubMed: 35022620]
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I (2017). Attention is all you need. Adv. Neural Inf. Process. Syst 30.
31. Rombach R, Blattmann A, Lorenz D, Esser P, and Ommer B High-resolution image synthesis with latent diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.
32. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. [PubMed: 34265844]
33. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876. [PubMed: 34282049]
34. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130. [PubMed: 36927031]
35. Gomes J, Ramsundar B, Feinberg EN, and Pande VS (2017). Atomic convolutional networks for predicting protein-ligand binding affinity. Preprint at arXiv.
36. Cunningham JM, Koytiger G, Sorger PK, and AlQuraishi M (2020). Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. Nat. Methods 17, 175–183. [PubMed: 31907444]
37. Torng W, and Altman RB (2019). High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics 35, 1503–1512. [PubMed: 31051039]
38. Corso G, Stärk H, Jing B, Barzilay R, and Jaakkola T (2023). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking The Eleventh International Conference on Learning Representations.
39. Kudo T, Meireles AM, Moncada R, Chen Y, Wu P, Gould J, Hu X, Kornfeld O, Jesudason R, Foo C, et al. (2024). Multiplexed, image-based pooled screens in primary cells and tissues with perturbview. Nat. Biotechnol, 1–10. [PubMed: 38191665]
40. Roohani Y, Huang K, and Leskovec J (2023). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat. Biotechnol 42, 927–935. [PubMed: 37592036]
41. Bunne C, Stark SG, Gut G, Del Castillo JS, Levesque M, Lehmann K-V, Pelkmans L, Krause A, and Ratsch G (2023). Learning single-cell perturbation responses using neural optimal transport. Nat. Methods 20, 1759–1768. [PubMed: 37770709]
42. Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, Srivatsan SR, Naghipourfar M, Daza RM, Martin B, et al. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. Mol. Syst. Biol 19, e11517. [PubMed: 37154091]
43. Bunne C, Krause A, and Cuturi M (2022). Supervised Training of Conditional Monge Maps. Adv. Neural Inf. Process. Syst 35, 6859–6872.
44. Bereket M, and Karaletsos T (2024). Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder. Adv. Neural Inf. Process. Syst 36.
45. Huang K, Huang K, Lopez R, Hutter J-C, Kudo T, Rios A, and Regev A (2023). Sequential Optimal Experimental Design of Perturbation Screens Guided by Multi-modal Priors. In International Conference on Research in Computational Molecular Biology (Springer Nature), pp. 17–37.
46. Roohani YH, Vora J, Huang Q, Liang P, and Leskovec J (2024). BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments Preprint at arXiv.
47. Cleary B, and Regev A (2024). The necessity and power of random, undersampled experiments in biology. Preprint at arXiv.
48. Ji Y, Zhou Z, Liu H, and Davuluri RV (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37, 2112–2120. [PubMed: 33538820]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

49. Brandes N, Ofer D, Peleg Y, Rappoport N, and Linial M (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38, 2102–2110. [PubMed: 35020807]
50. Celaj A, Gao AJ, Lau TTY, Holgersen EM, Lo A, Lodaya V, Cole CB, Denroche RE, Spickett C, Wagih O, et al. (2023). An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. Preprint at bioRxiv.
51. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science 386, eado9336. [PubMed: 39541441]
52. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, et al. (2024). Simulating 500 million years of evolution with a language model. Preprint at bioRxiv.
53. Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, Hussain R, Russ J, Yip J, Hill E, Pacesa M, et al. (2024). Design of highly functional genome editors by modeling the universe of CRISPR-cas sequences. Preprint at bioRxiv.
54. Peng Z, Schussheim B, and Chatterjee P (2024). PTM-mamba: a PTM-aware protein language model with bidirectional gated mamba blocks. Preprint at bioRxiv.
55. Dai B, Mattox DE, and Bailey-Kellogg C (2021). Attention please: modeling global and local context in glycan structure-function relationships. Preprint at bioRxiv.
56. Yu T, Yao T, Sun Z, Shi F, Zhang L, Lyu K, Xuan B, Liu A, Zhang X, Zou J, et al. (2024). LipidBERT: A Lipid Language Model Pretrained on METiS de novo Lipid Library. Preprint at arXiv.
57. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, et al. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 384, eadl2528. [PubMed: 38452047]
58. Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, and Leskovec J (2024). Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with Saturn. Nat. Methods 21, 1492–1500. [PubMed: 38366243]
59. Rosen Y, Roohani Y, Agrawal A, Samotorčan L, Tabula Sapiens Consortium, Quake SR, and Leskovec J (2023). Universal cell embeddings: A foundation model for cell biology. Preprint at bioRxiv.
60. Chen Y, and Zou J (2024). GenePT: A Simple but Effective Foundation Model for Genes and Cells Built from ChatGPT. Preprint at bioRxiv.
61. Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, Zhang C, Le T, Johansson F, Schutten R, et al. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature 590, 649–654. [PubMed: 33627808]
62. Chandrasekaran SN, Cimini BA, Goodale A, Miller L, Kost-Alimova M, Jamali N, Doench JG, Fritchman B, Skepner A, Melanson M, et al. (2024). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nat. Methods 21, 1114–1121. [PubMed: 38594452]
63. Carlson RJ, Leiken MD, Guna A, Hacohen N, and Blainey PC (2023). A genome-wide optical pooled screen reveals regulators of cellular antiviral responses. Proc. Natl. Acad. Sci. USA 120, e2210623120. [PubMed: 37043539]
64. Thul PJ,Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al. (2017). A subcellular map of the human proteome. Science 356, eaal3321. [PubMed: 28495876]
65. McDole K, Guignard L, Amat F, Berger A, Malandain G, Royer LA, Turaga SC, Branson K, and Keller PJ (2018). In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. Cell 175, 859–876.e33. [PubMed: 30318151]
66. Nogales E, and Mahamid J (2024). Bridging structural and cell biology with cryo-electron microscopy. Nature 628, 47–56. [PubMed: 38570716]
67. Bauda E, Gallet B, Moravcova J, Effantin G, Chan H, Novacek J, Jouneau PH, Rodrigues CDA, Schoehn G, Moriscot C, et al. (2024). Ultrastructure of macromolecular assemblies contributing to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bacterial spore resistance revealed by in situ cryo-electron tomography. Nat. Commun 15, 1376. [PubMed: 38355696]

68. Lelek M, Gyparaki MT, Beliu G, Schueder F, Griffié J, Manley S, Jungmann R, Sauer M, Lakadamyali M, and Zimmer C (2021). Single-molecule localization microscopy. Nat. Rev. Methods Primers 1, 39. [PubMed: 35663461]
69. Möckl L, and Moerner WE (2020). Super-resolution microscopy with single molecules in biology and beyond–essentials, current trends, and future challenges. J. Am. Chem. Soc 142, 17828–17844. [PubMed: 33034452]
70. Cesnik A, Schaffer LV, Gaur I, Jain M, Ideker T, and Lundberg E (2024). Mapping the multiscale proteomic Organization of cellular and Disease Phenotypes. Annu. Rev. Biomed. Data Sci 7, 369–389. [PubMed: 38748859]
71. Qin Y, Huttlin EL, Winsnes CF, Gosztyla ML, Wacheul L, Kelly MR, Blue SM, Zheng F, Chen M, Schaffer LV, et al. (2021). A multi-scale map of cell structure fusing protein images and interactions. Nature 600, 536–542. [PubMed: 34819669]
72. Dosovitskiy A (2020). An image is worth 16×16 words: transformers for image recognition at scale. Preprint at arXiv.
73. Fukushima K (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern 36, 193–202. [PubMed: 7370364]
74. LeCun Y, and Yoshua B (1995). Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks 3361, 255–258.
75. Bao Y, Sivanandan S, and Karaletsos T (2023). Channel Vision Transformers: An Image Is Worth c × 16 × 16 WordsThe Twelfth. International Conference on Learning Representations 4.
76. Kraus O, Kenyon-Dean K, Saberian S, Fallah M, McLean P, Leung J, Sharma V, Khan A, Balakrishnan J, Celik S, et al. (2024). Masked autoencoders for microscopy are scalable learners of cellular biology. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11757–11768.
77. Bao Y, and Karaletsos T (2023). Contextual vision transformers for robust representation learning. Preprint at arXiv.
78. Lopez R, Regier J, Cole MB, Jordan MI, and Yosef N (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods 15, 1053–1058. [PubMed: 30504886]
79. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. (2023). Transfer learning enables predictions in network biology. Nature 618, 616–624. [PubMed: 37258680]
80. Kobayashi-Kirschvink KJ, Comiter CS, Gaddam S, Joren T, Grody EI, Ounadjela JR, Zhang K, Ge B, Kang JW, Xavier RJ, et al. (2024). Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA. Nat. Biotechnol 42, 1726–1734. [PubMed: 38200118]
81. Ryu J, Lopez R, Bunne C, and Regev A (2024). Cross-modality matching and prediction of perturbation responses with labeled Gromov-Wasserstein optimal transport. Preprint at arXiv.
82. Saar KL, Scrutton RM, Bloznelyte K, Morgunov AS, Good LL, Lee AA, Teichmann SA, and Knowles TPJ (2024). Protein Condensate Atlas from predictive models of heteromolecular condensate composition. Nat. Commun 15, 5418. [PubMed: 38987300]
83. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas A, Kamitaki N, Martersteck E, et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214. [PubMed: 26000488]
84. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 78–82. [PubMed: 27365449]
85. Lundberg E, and Borner GHH (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol 20, 285–302. [PubMed: 30659282]
86. Marconato L, Palla G, Yamauchi KA, Virshup I, Heidari E, Treis T, Vierdag WM, Toth M, Stockhaus S, Shrestha RB, et al. (2024). SpatialData: an open and universal data framework for spatial omics. Nat. Methods 10.1038/s41592-024-02212-x.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

87. Somnath VR, Pariset M, Hsieh Y-P, Martinez MR, Krause A, and Bunne C (2023). Aligned Diffusion Schrödinger Bridges. In Uncertainty in Artificial Intelligence, pp. 1985–1995.

88. Katharopoulos A, Vyas A, Pappas N, and Fleuret F (2020). Fast autoregressive transformers with linear attention. In International Conference on Machine Learning.

89. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500. [PubMed: 38718835]

90. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, Gilbert LA, and Weissman JS (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science 365, 786–793. [PubMed: 31395745]

91. Lawson MJ, Camsund D, Larsson J, Baltekin Ö, Fange D, and Elf J (2017). In situ genotyping of a pooled strain library after characterizing complex phenotypes. Mol. Syst. Biol 13, 947. [PubMed: 29042431]

92. Papamarkou T, Skoularidou M, Palla K, Aitchison L, Arbel J, Dunson D, Filliponne M, Fortuin V, Hennig P, Hernandez-Lobato JM, et al. (2024). Position: bayesian deep learning is needed in the age of large-scale AI. In Forty-First International Conference on Machine Learning.

93. D'Angelo F, and Fortuin V (2021). Wenzel F.On Stein Variational Neural Network. Ensembles Preprint at arXiv.

94. Ober SW, Rasmussen CE, and van der Wilk M (2021). The promises and pitfalls of deep kernel learning. In Conference on Uncertainty in Artificial Intelligence, pp. 1206–1216.

95. Karaletsos T (2020). Bui T.D.Hierarchical Gaussian Process Priors for Bayesian Neural Network Weights. Adv. Neural Inf. Process. Syst 33, 17141–17152.

96. Kapoor S, Maddox WJ, Izmailov P, and Wilson AG (2022). On uncertainty, tempering, and data augmentation in bayesian classification. Adv. Neural Inf. Process. Syst 35, 18211–18225.

97. Lakshminarayanan B, Pritzel A, and Blundell C (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Adv. Neural Inf. Process. Syst 30.

98. Angelopoulos AN, and Bates S (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Preprint at arXiv.

99. Cherian JJ, Gibbs I, and Candès EJ (2024). Large language model validity via enhanced conformal prediction methods. Preprint at arXiv.

100. Cho NH, Cheveralls KC, Brunner AD, Kim K, Michaelis AC, Raghavan P, Kobayashi H, Savy L, Li JY, Canaj H, et al. (2022). Open-Cell: endogenous tagging for the cartography of human cellular organization. Science 375, eabi6983. [PubMed: 35271311]

101. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. (2015). Proteomics. Tissue-based map of the human proteome. Science 347, 1260419. [PubMed: 25613900]

102. Berger C, Premaraj N, Ravelli RBG, Knoops K, López-Iglesias C, and Peters PJ (2023). Cryo-electron tomography on focused ion beam lamellae transforms structural cell biology. Nat. Methods 20, 499–511. [PubMed: 36914814]

103. Loconte V, Chen JH, Vanslembrouck B, Ekman AA, McDermott G, Le Gros MA, and Larabell CA (2023). Soft X-ray tomograms provide a structural basis for whole-cell modeling. FASEB J. 37, e22681. [PubMed: 36519968]

104. Moffitt JR, Lundberg E, and Heyn H (2022). The emerging landscape of spatial profiling technologies. Nat. Rev. Genet 23, 741–759. [PubMed: 35859028]

105. Vandereyken K, Sifrim A, Thienpont B, and Voet T (2023). Methods and applications for single-cell and spatial multi-omics. Nat. Rev. Genet 24, 494–515. [PubMed: 36864178]

106. Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaup B, Brown P, et al. (2022). The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science 376, eabl4896. [PubMed: 35549404]

107. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, and Zou J (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng 4, 827–834. [PubMed: 32572199]

108. Bock C, Boutros M, Camp JG, Clarke L, Clevers H, Knoblich JA, Liberali P, Regev A, Rios AC, Stegle O, et al. (2021). The organoid cell atlas. Nat. Biotechnol 39, 13–17. [PubMed: 33384458]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

109. Tabula; Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. Nature 562, 367–372. [PubMed: 30283141]
110. Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbić M, Spanier K, et al. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. Science 375, eabk2432. [PubMed: 35239393]
111. Lange M, Granados A, Vijaykumar S, Bragantini J, Ancheta S, Santhosh S, Borja M, Kobayashi H, McGeever E, Solak AC, et al. (2023). Zebrahub – Multimodal zebrafish Developmental Atlas Reveals the State Transition Dynamics of Late Vertebrate Pluripotent Axial Progenitors. Preprint at bioRxiv.
112. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, and O'Sullivan C (2022). The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 50, D387–D390. [PubMed: 34850094]
113. Achiam J, et al. (2023). GPT-4 technical report. Preprint at arXiv.
114. Ding F, and Steinhardt JN (2024). Protein language models are biased by unequal sequence sampling across the tree of life. Preprint at bioRxiv.
115. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. (2023). A draft human pangenome reference. Nature 617, 312–324. [PubMed: 37165242]
116. Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H, and Cui P (2021). Towards out-of-distribution generalization: A survey. Preprint at arXiv.
117. Nisonoff H, Wang Y, and Listgarten J (2023). Coherent blending of biophysics-based knowledge with bayesian neural networks for robust protein property prediction. ACS Synth. Biol 12, 3242–3251. 10.1021/acssynbio.3c00217. [PubMed: 37888887]
118. Zheng F, Kelly MR, Ramms DJ, Heintschel ML, Tao K, Tutuncuoglu B, Lee JJ, Ono K, Foussard H, Chen M, et al. (2021). Interpretation of cancer mutations using a multiscale map of protein systems. Science 374, eabf3067. [PubMed: 34591613]
119. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, and Ideker T (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods 15, 290–298. [PubMed: 29505029]
120. Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, Ektefaie Y, Kondic J, and Zitnik M (2024). Empowering biomedical discovery with AI agents. Cell 187, 6125–6151. [PubMed: 39486399]
121. Hurrell T, Naidoo J, Ntlhafu T, and Scholefield J (2024). An African perspective on genetically diverse human induced pluripotent stem cell lines. Nat. Commun 15, 8581. [PubMed: 39362853]
122. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, Ch'ang LY, Huang W, Shen B, Tam Y, et al. (2003). The international HapMap project. Nature 5, 467–475.
123. Cancer; Genome; Atlas; Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, and Stuart JM (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet 45, 1113–1120. [PubMed: 24071849]
124. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]
125. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, and Foster B (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet 45, 580–585. [PubMed: 23715323]
126. Pontén F, Jirström K, and Uhlen M (2008). The Human Protein Atlas–a tool for pathology. J. Pathol 216, 387–393. [PubMed: 18853439]
127. Downey P, and Peakman TC (2008). Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles. Int. J. Epidemiol 37 (Suppl 1), i46–i50. [PubMed: 18381393]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

128. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Xiong C, Sun ZZ, Socher R, et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol 41, 1099–1106. [PubMed: 36702895]

129. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet 47, 856–860. [PubMed: 26121088]

130. Mason C, Brindley DA, Culme-Seymour EJ, and Davie NL (2011). Cell therapy industry: billion dollar global business with unlimited potential. Regen. Med 6, 265–272. [PubMed: 21548728]

131. Bashor CJ, Hilton IB, Bandukwala H, Smith DM, and Veiseh O (2022). Engineering the next generation of cell-based therapeutics. Nat. Rev. Drug Discov 21, 655–675. [PubMed: 35637318]

132. Jia Q, Wang A, Yuan Y, Zhu B, and Long H (2022). Heterogeneity of the tumor immune microenvironment and its clinical relevance. Exp. Hematol. Oncol 11, 24. [PubMed: 35461288]

133. Melssen MM, Sheybani ND, Leick KM, and Slingluff CL (2023). Barriers to immune cell infiltration in tumors. J. Immunother. Cancer 11.

134. Chow A, Perica K, Klebanoff CA, and Wolchok JD (2022). Clinical implications of T cell exhaustion for cancer immunotherapy. Nature Reviews Clinical Oncology 19, 775–790.

135. de Visser KE, and Joyce JA (2023). The evolving tumor microenvironment: from cancer initiation to metastatic outgrowth. Cancer Cell 41, 374–403. [PubMed: 36917948]

136. Barkley D, Moncada R, Pour M, Liberman DA, Dryg I, Werba G, Wang W, Baron M, Rao A, Xia B, et al. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat. Genet 54, 1192–1201. [PubMed: 35931863]

137. Schwartzberg L, Kim ES, Liu D, and Schrag D (2017). Precision oncology: who, how, what, when, and when not? American Society of Clinical Oncology Educational Book 37, 160–169. [PubMed: 28561651]

138. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, et al. (2018). How many human proteoforms are there? Nat. Chem. Biol 14, 206–214. [PubMed: 29443976]

139. Katsoulakis E, Wang Q, Wu H, Shahriyari L, Fletcher R, Liu J, Achenie L, Liu H, Jackson P, Xiao Y, et al. (2024). Digital twins for health: a scoping review. npj Digit. Med 7, 77. [PubMed: 38519626]

140. Rajewsky N, Almouzni G, Gorski SA, Aerts S, Amit I, Bertero MG, Bock C, Bredenoord AL, Cavalli G, Chiocca S, et al. (2020). Lifetime and improving European healthcare through cell-based interceptive medicine. Nature 587, 377–386. [PubMed: 32894860]

141. Alix-Panabières C, and Pantel K (2021). Liquid biopsy: from discovery to clinical application. Cancer Discov. 11, 858–873. [PubMed: 33811121]

142. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, and Regev A (2022). The evolution, evolvability and engineering of gene regulatory DNA. Nature 603, 455–463. [PubMed: 35264797]

143. Gómez-de-Mariscal E, García-López-de-Haro C, Ouyang W, Donati L, Lundberg E, Unser M, Muñoz-Barrutia A, and Sage D (2021). DeepImageJ: A user-friendly environment to run deep learning models in ImageJ. Nat. Methods 18, 1192–1195. [PubMed: 34594030]

144. Le T, Winsnes CF, Axelsson U, Xu H, Mohanakrishnan Kaimal J, Mahdessian D, Dai S, Makarov IS, Ostankovich V, Xu Y, et al. (2022). Analysis of the human protein atlas weakly supervised single-cell classification competition. Nat. Methods 19, 1221–1229. [PubMed: 36175767]

145. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, Chen B, Zhang A, Shao D, Shaban M, et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med 30, 850–862. [PubMed: 38504018]

146. Moen E, Bannon D, Kudo T, Graf W, Covert M, and Van Valen D (2019). Deep learning for cellular image analysis. Nat. Methods 16, 1233–1246. [PubMed: 31133758]

147. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet 53, 354–366. [PubMed: 33603233]

148. Ho J, Jain A, and Abbeel P (2020). Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst 33, 6840–6851.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

149. Lipman Y, Chen RT, Ben-Hamu H, Nickel M, and Le M (2023). Flow Matching for Generative Modeling. International Conference on Learning Representations.

150. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, and Monfardini G (2009). The graph neural network model. IEEE Trans. Neural Netw 20, 61–80. [PubMed: 19068426]

151. Cao Y, and Shen Y (2020). Energy-based graph convolutional networks for scoring protein docking models. Proteins 88, 1091–1099. [PubMed: 32144844]

152. Brbić M, Cao K, Hickey JW, Tan Y, Snyder MP, Nolan GP, and Leskovec J (2022). Annotation of spatially resolved single-cell data with STELLAR. Nat. Methods 19, 1411–1418. [PubMed: 36280720]

153. Wu Z, Trevino AE, Wu E, Swanson K, Kim HJ, D'Angio HB, Preska R, Charville GW, Dalerba PD, Egloff AM, et al. (2022). Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. Nat. Biomed. Eng 6, 1435–1448. [PubMed: 36357512]

154. Hamilton W, Ying Z, and Leskovec J (2017). Inductive Representation Learning on Large Graphs. Adv. Neural Inf. Process. Syst 30.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 1.

Grand challenges for building the AIVC

OUTLINING CAPABILITIES AND DESIGNING EVALUATION FRAMEWORKS

The burgeoning number of foundation models in biology perform a subset of the capabilities of virtual cells outlined in this perspective. Given the diversity of these approaches, it is important to define what the core capabilities of AIVCs should be and how those capabilities can be evaluated. For every capability, proper metrics must be designed, and comprehensive evaluation data be collected. Models' capabilities should be assessed on general performance as well as on their ability to answer specific biological questions. It is imperative to continuously improve benchmarking strategies along with AIVC models and ensure that they align with biologically meaningful objectives. As the field develops better alignment on these questions, collaborative opportunities will arise, and the speed at which virtual cells can be generated will accelerate.

ESTABLISHING SELF-CONSISTENCY ACROSS VARYING CONTEXTS WITH DIFFERENT ARCHITECTURES

Biology is tremendously complex: it operates across different scales, in different contexts, and is measured with different modalities. AIVC models must be self-consistent across all of these axes. Models should propagate function across physical scales—interactions between molecules should have consistent effects when measuring binding affinity, gene expression, cell-cell communication, or tissue organization. As physical and dynamic scales increase in scope and size, additional context, for example, species, cell type, tissue, disease status, etc., should fine-tune predictions made at smaller resolutions, while also accounting for stochasticity. Model predictions should also be agnostic to their input and output modalities. The same entity, profiled with different technologies, should have the same internal representation in an AIVC. To properly model such complex behaviors, many machine learning approaches should be explored and their merits carefully judged.

BALANCING INTERPRETABILITY AND BIOLOGICAL UTILITY

A consistent trend in the application of deep learning methods to biology, accelerated by the rise of large foundation models, has been the implicit trade-off between models' performance gains and their increasingly uninterpretable "black box" natures. AIVC models will ultimately be judged on their ability to expand our understanding of biology, either by providing novel insights to biological processes or by accelerating the scientific process. To achieve this goal, AIVC models must make highly accurate and well-calibrated predictions that simulate biology, and the trade-off between actionability and interpretability will have to be balanced. Actionable model outputs are those of high utility to design affordable and efficient validation experiments and are key for initial real-world use. Various approaches exist for explaining model predictions, including causal modeling, sparse featurization, and counterfactual reasoning, and this is a highly active research area. Building intuitive interfaces that facilitate the study

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and interpretation of AIVCs via other models, such as AI research agents, will further increase downstream utility.

CONSTRUCTING A FRAMEWORK FOR COLLABORATIVE CELL MODELING

The successful development of AIVCs will require collaboration across disciplines. We foresee a future where AIVC platforms function as open, interconnected hubs for collaborative development and broad deployment of cell models to researchers and as education hubs delivering training to researchers, as well as providing engagement activities for educators, patients, and the public. Thus, investments in infrastructure fostering open and collaborative development of AIVCs should be of high priority.

ENSURING AIVCs BENEFIT ALL AND PROMOTE ETHICAL AND RESPONSIBLE USE

Generating large open datasets that reflect human diversity—datasets integral for training AIVC models—poses a substantial challenge. Developers will have to use the utmost care to ensure these datasets are used ethically and transparently while building AIVCs and develop strategies to mitigate risks of model contamination with falsified data. Early adopters of AIVCs will have a key role in promoting and demonstrating responsible use of these models. Furthermore, the development of chat-based interfaces could be crucial in democratizing access to AIVCs. Close collaboration with ethics and regulatory experts from the outset is paramount for establishing new regulatory norms that will facilitate the responsible use of AIVCs.

UNDERSTANDING THE VALUE OF DIFFERENT DATA TYPES TO PRIORITIZE LARGE-SCALE DATA GENERATION

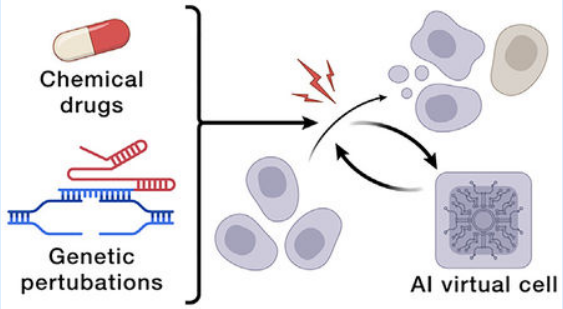
A fundamental question for the collaborative development of AIVCs is which data and modalities should be collected to enable generalization across biological contexts and scales. These data will need to encompass the breadth of biology in different species, domains, and modalities, representing the heterogeneity of life, while maintaining depth sufficient to distinguish true signals from noise. A key aspect of data generation will be the simultaneous measurement of temporal and physical scales, while also allowing perturbations of the system.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 2.

Vignettes

CELL ENGINEERING TO ENABLE PHENOTYPIC DRUG DISCOVERY AND CELL-BASED THERAPEUTICS

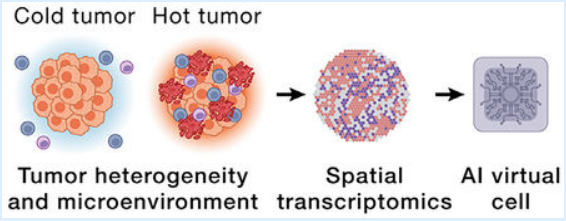


One challenge in developing successful therapies is the difficulty in incorporating the full underlying genetic, molecular and cellular basis of disease during drug discovery and development.[129] These context-specific underpinnings are not fully specified and often vary between human patients and model systems used in pre-clinical studies. By integrating biological data from various sources relevant to specific disease contexts, the AIVC could generate an environment for testing different therapeutic interventions *in silico* and identify approaches for engineering cells to reverse disease phenotypes, while accounting for the effects of varying both treatments and patient profiles. By representing the overall disease phenotype specific to patient populations (rather than one specific biochemical target at a time), the AIVC can enable virtual phenotypic screens. Although *in silico* experiments may not always be fully accurate, by prioritizing virtual hits with higher chances of success, the AIVC can lower experimentation costs and accelerate the process.

The AIVC has potential to push the cell therapy frontier. With growing evidence affirming the efficacy and safety of cell-based therapies for rare diseases and cancer,[130,131] the AIVC can improve systematization and precision to cell engineering. For example, virtual cell-based engineering could enable targeted modifications to pancreatic beta cells to create individualized beta cell replacement therapies for type 1 diabetes. By simulating the biological phenotype of individual patients, *in silico* experiments within the AIVC could identify interventions that help drive the differentiation of beta cells from progenitors, cloak them from the immune system, and maintain their function, with the ultimate goal of either transplanting these engineered cells into patients or engineering them *in situ*.

UNLOCKING THE POWER OF SPATIAL BIOLOGY TO FIGHT CANCER

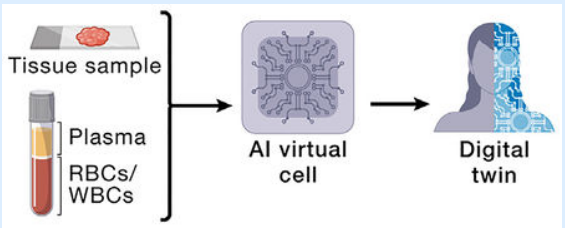
Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Spatial structures in cancer, specifically within the tumor microenvironment (TME), are critical drivers of cancer progression and can drive resistance to the immune system and limit drug efficacy.[132] Malignant cells within a tumor can engage in active immune evasion by either blocking immune infiltration,[133] evading immune recognition, or dampening immune cell function.[134] Thus, immune resistance must be understood in the spatial context of the cellular neighborhood to identify the specific cell states and gene signatures involved. Although next-generation spatial profiling methods enable researchers to experimentally investigate the heterogeneity of the TME,[135] an AIVC could extend these analyses to a universal, pan-cancer framework, which can be personalized to individual patients. Using an AIVC model, cancer researchers should be able to identify TME niches shared across multiple cancer types from many patients. Identifying pan-cancer markers can drive cancer treatment both by highlighting new targets and also by identifying existing treatments that can be applied to new cancer types.[136] In this setting, the AIVC would help identify the interactions associated with TME cell states and would search for similar states from any disease where existing treatments exist.

Finally, the AIVC could greatly advance precision oncology.[137] Given that the AIVC will capture intrinsic variation, the genetic diversity of individual patient's cancers will be represented in any analyses. Although the AIVC would already accurately qualify the change in the expression of genes, tumor sequencing data would allow it to model the change of function of those genes, for example, through loss of function, change in post-translational modifications, or rewiring of protein-protein interactions and signaling networks.[138]

DIAGNOSTIC VIRTUAL CELL MODELS FOR INDIVIDUAL PATIENTS



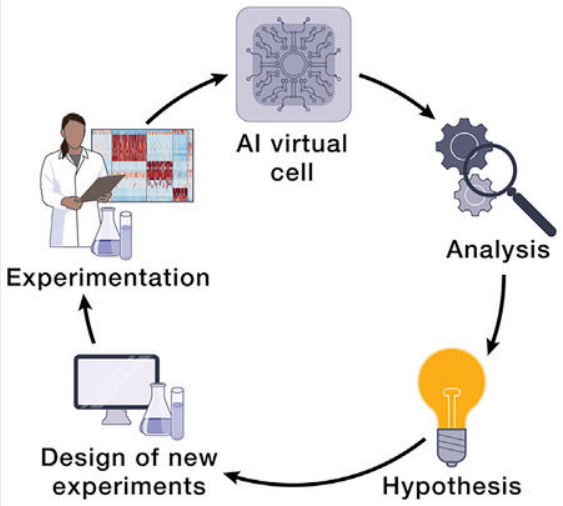
The AIVC could introduce a new approach to diagnostics that incorporates a personalized AIVC (or a digital twin[139]) to track a patient's health and suggest suitable interventions. The AIVC would create a detailed representation of each patient's cells by incorporating specific patient data, such as genetic sequences, single-cell profiles from blood, and tissue pathology images, along with additional clinical information from their health records. Periodic updates to each patient's AIVC instance enable monitoring

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of evolving health conditions, prediction of upcoming adverse events, and potential therapeutic outcomes.

Through additional updates from less costly assays, this virtual patient model could be progressively refined and made more robust.[140] For example, transcriptomic or genetic liquid biopsies can reveal significant and diverse characteristics of a patient from a single test and could greatly aid in the diagnosis of a broad spectrum of conditions.[141] Through the virtual cell's implicit and structured representation of universal cell types and states, one can envision the creation of patient models of inaccessible cell types, such as beta cells in the pancreas or neurons in the brain, generated after sampling accessible cell types such as blood or skin.

A HYPOTHESIS-GENERATING FRAMEWORK FOR SCIENTIFIC RESEARCH



Traditionally, the biological research community has relied on computational models for analyzing data from past experiments based on an existing hypothesis. The virtual cell could switch the paradigm by computationally exploring a vast array of possible hypotheses through *in silico* experimentation. It could identify the most informative experiments for addressing specific biological questions, shifting the role of computational models from merely validating hypotheses or processing observations without a particular goal to generating specific sets of hypotheses to pursue.

This shift could greatly enhance the scientific discovery process: instead of conducting a single experiment followed by an in-depth analysis, scientists can engage in a dynamic iterative interaction with the virtual cell. With each new piece of data, they can refine their understanding of the biological system and consult the virtual cell to identify what additional experimental data could be valuable. Ultimately, we may be able to perform active learning with biologists in the loop and construct self-driving labs for efficient and unbiased generation of virtual cells.

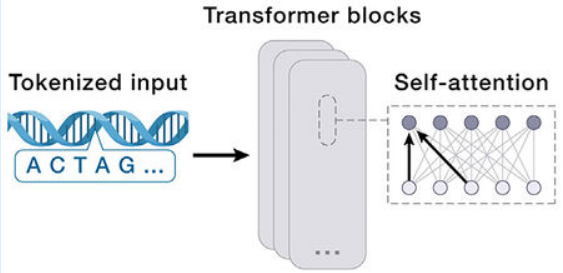
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 3.

AI techniques for building the AIVC

The AIVC will connect a number of diverse neural network architectures. Although these architectures may not have been purpose-built for biological applications, they have each demonstrated success when matched with specific biological modalities and inductive biases. In many cases, these architectures may be exchangeable, and one must weigh their individual trade-offs in accuracy, speed, and generalizability. Beyond this, the community is actively developing AI architectures tailored to the characteristics of (large) biological datasets.

TRANSFORMERS



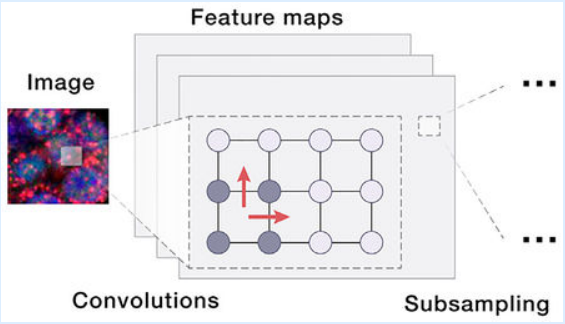
A transformer neural network[30] comprises multiple transformer layers, each taking a series of tokens (discrete pieces of information such as words, RNA molecules, or gene representations) as input—initial tokens for the first layer and outputs from the preceding layer for subsequent ones. Within each layer, tokens use self-attention to integrate context from other tokens, enhancing their own representations, which are then processed through a feed forward network. This architecture, which fundamentally requires only a collection of tokens, adapts well across various applications and use cases.

The collection of tokens passed to a transformer does not have any ordering by default. Additionally, the self-attention mechanism, the core of the success of the transformer, can be taken as a strong biological inductive bias. For instance, in representing cells through their RNA molecules detected via scRNA-seq, each RNA molecule, represented as a token, interacts with others, modeling gene interactions through self-attention.[30] Customizing input tokens with numerical representations of genes further allows the integration of diverse biological data scales, from individual genes to whole cells.[59,60]

Additionally, introducing positional encodings to tokens enables transformers to process sequences, such as natural language,[30] or biological sequences, such as DNA,[48,142] by incorporating sequence-specific dependencies. This approach is crucial in applications such as masked language modeling, where the model predicts missing tokens in sequences, enhancing its understanding of contextual relationships within data. Innovations continue to refine transformers, increasing their capacity to handle longer sequences and improving efficiency, with advancements such as state-space models enabling the generation of extensive DNA sequences.[51]

CNN

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



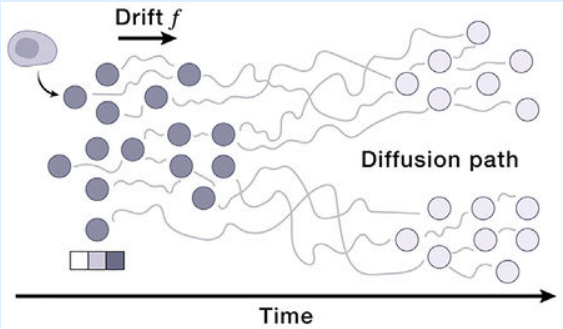
A CNN is a deep learning model primarily used for analyzing images.[73,74] It consists of multiple layers that automatically and adaptively learn spatial hierarchies of features through backpropagation. This learning is facilitated by convolutional layers that apply filters to local patches of input data, pooling layers that reduce dimensionality, and fully connected layers that interpret the features extracted to make decisions.

In the field of biology, CNNs have proven invaluable for tasks involving image data due to their ability to detect complex patterns and structures, such as microscope images of cells and tissues. Here, CNNs play a critical role in multiplex imaging,[143] where multiple targets within a single sample are labeled and visualized simultaneously. This technique is particularly useful in studying the complex interactions of different molecules or cell types within a heterogeneous tissue environment.[144] Another notable application is in the analysis of H&E-stained tissue sections, commonly used in clinical pathology.[145] Lastly, in live-cell imaging, CNNs are employed to track dynamic changes within cells or even single molecules over time, providing insights into cell migration, responses to treatment, or the movement and interaction of individual molecules within cells, revealing crucial biological processes at a molecular level.[146]

Beyond their traditional use in image processing, CNNs can also be applied to model sequence data, such as DNA sequences, where they identify patterns and features that are predictive of biological functions.[147] Despite their extensive utility, CNNs are increasingly being supplemented or replaced by vision transformer models,[72] which leverage self-attention mechanisms to process entire images in parallel. These models can often achieve higher accuracy on tasks where understanding the global context within the image is crucial.

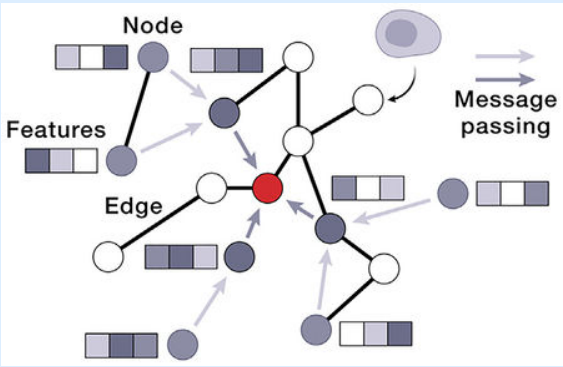
DIFFUSION MODELS

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Diffusion models are a class of generative deep learning models that have recently gained attention for their ability to generate high-quality, diverse samples across various domains.[148] They operate by gradually transforming a distribution of random noise into a structured output (images, text, cellular states, etc.) through a process that mimics a physical diffusion process. Building up on diffusion model architectures, approaches such as flow matching methods can also model the distributional evolution over time,[149] making them especially powerful in biological applications where dynamic changes and temporal progression are critical. Flow matching methods thus capture and generate sequences of data that reflect continuous transformations, such as the developmental stages of cells over time and space or the response of biological systems to treatments.[87] The ability of diffusion and flow matching models to learn and replicate complex distributions, combined with the temporal and spatial modeling capabilities of flow matching methods, makes them particularly suited for tasks that involve high-dimensional, intricate data structures typical of biological systems.

GNNs



GNNs are a set of architectures that can model graphical data.[150] Graphs, sets of nodes connected by edges, are useful representations for many kinds of biological data. When modeling a biological system, a GNN could be a good choice if a graph structure represents some core inductive bias. For example, a protein structure[151] can be thought of as a graph where residues are nodes, and their bonds are edges. Cells in a tissue form a graph: each cell is a node, and the cells it is physically proximal to are connected by edges.[152,153] In both cases, the graph represents how nodes are physically proximal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to each other. For spatially organized cells, the graph represents how they may pass chemical signals between one another.

GNNs can be used to make predictions about individual nodes, edges, or the graph as a whole.[154] For simplicity, in the following section, we describe a node-based GNN. At each layer, a node updates its representation using a neural network, which can take in that node's current representation, in addition to the representations of the node's neighbors, which are connected by an edge. By stacking GNN layers, a node can receive "messages" from neighboring nodes at increasing distances, "hops," from it. Nodes and edges can both be initialized with different features, which control their final representation and what messages they pass to their neighbors. For example, a GNN trained on spatial transcriptomic data could take node features to be the virtual cell representation of each cell's gene expression. The GNN would then update those representations to include context about each cell's neighbors, helping to identify spatial interactions and niches.[153]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

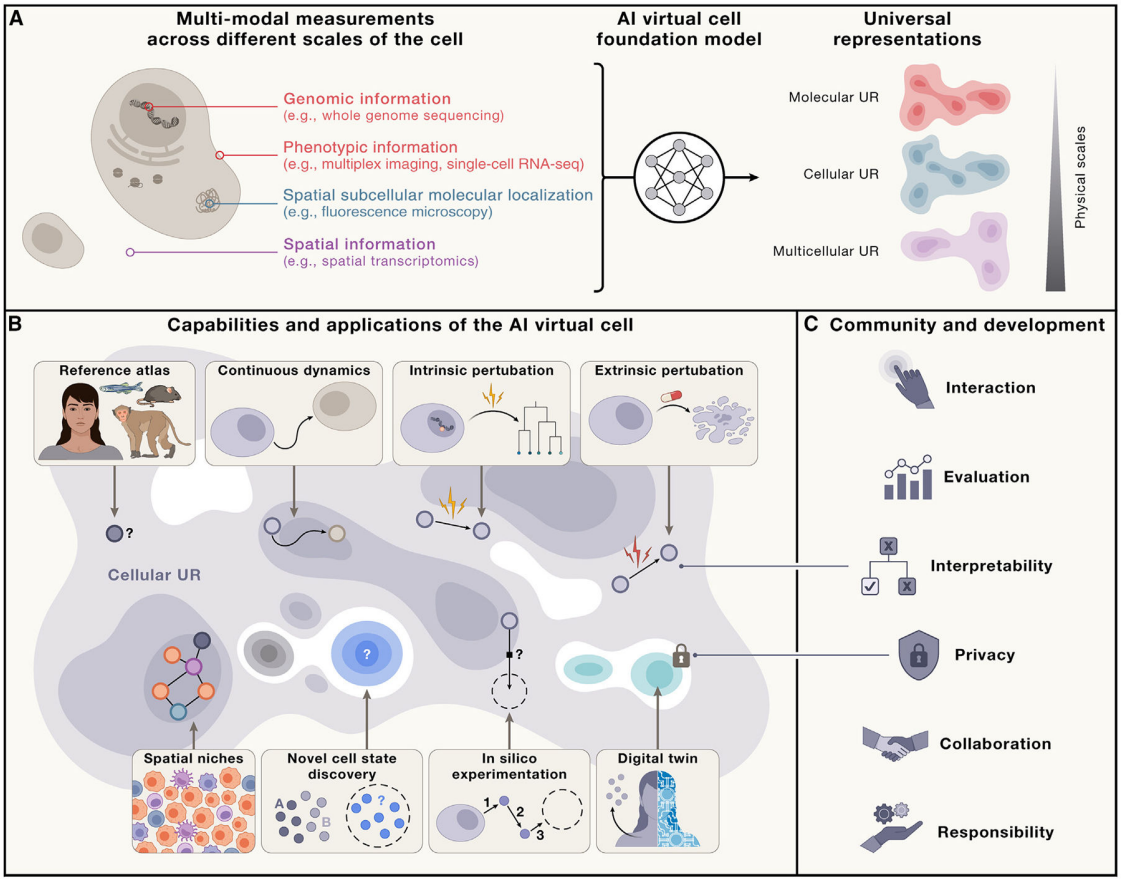


Figure 1. Capabilities of the AIVC

(A) The AIVC provides a universal representation (UR) of a cell state that can be obtained across species and conditions and generated from different data modalities across scales (molecular, cellular, and multicellular).

(B) The AIVC possesses capabilities to represent and predict cell biology. This universality allows the representation to act as a reference that can generalize to previously unobserved cell states, providing guidance for future data generation. Because the representation is shared across modalities, it also remains invariant to the specific data type used to generate it, serving as a virtual representation for unified analysis across modalities. The AIVC also allows modeling the dynamics of cells as they transition between different states, whether naturally due to processes such as differentiation or due to genetic variation or artificially through engineered perturbations. Thus, the AIVC enables *in silico* experimentation that would otherwise be cost-prohibitive or impossible in a lab.

(C) The utility of the AIVC depends on its interactions with humans at different levels. At the individual scientist level, it must be accessible through open licenses and the democratization of computing resources. Interpretability can be established through intermediary layers, such as language models that allow the virtual cell to communicate its results effectively. At the scientific community level, evaluating the AIVC should focus on core capabilities that move beyond narrow benchmarks. Community development will be

crucial for ongoing improvements to the virtual cell that remain accessible. At the societal level, the AIVC must ensure the privacy of its contents to protect sensitive data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

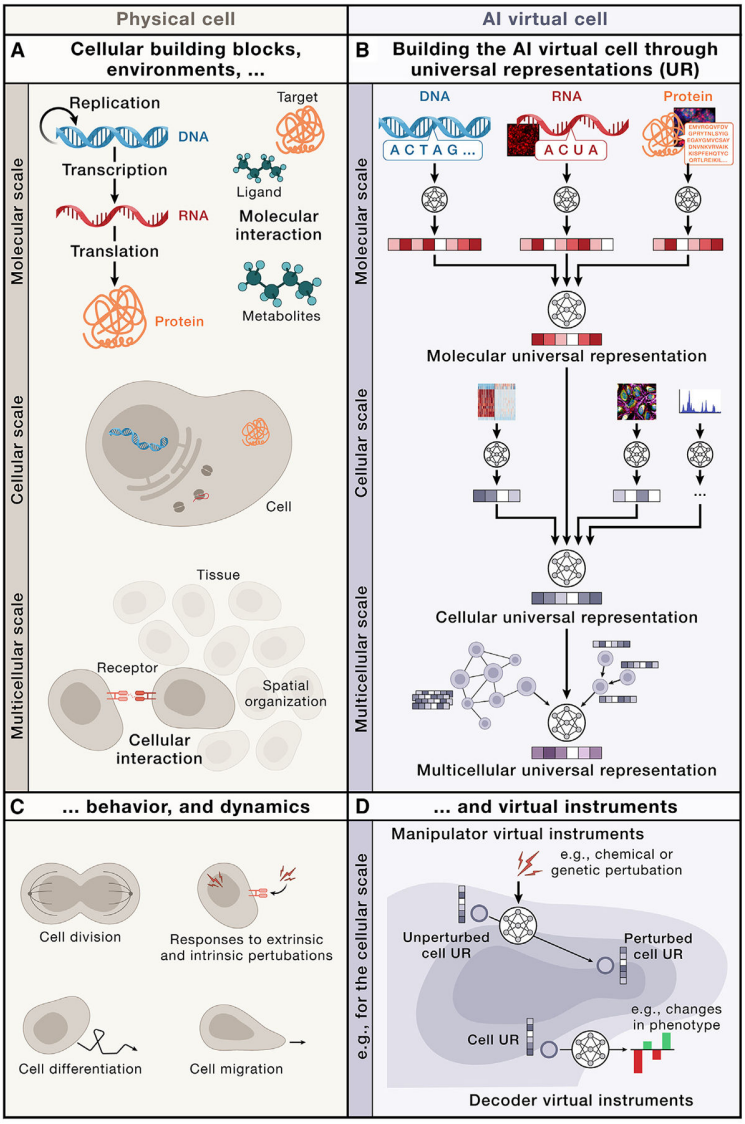


Figure 2. Overview of the AIVC

(A and B) (A) Similar to biological cells, (B) the AIVC models cell biology across different physical scales, including molecular, cellular, and multicellular. Along the physical dimension, the first scale models the state and interactions of individual molecules, such as those of the central dogma, as well as additional molecules, such as metabolites. Molecules can be represented as sequences or atomic structures. The next scale represents cells as collections of these molecules. For example, such cells contain a genetic sequence, RNA transcripts, and some quantities of proteins. Molecules within cells have specific locations that may be related to their function. The final scale models the interactions between cells and how they communicate and form complex tissues. Each scale relies on universal representations that are learned from multi-modal data and are integrating URs from the previous scale.

(C and D) (C) To capture the behavior and dynamics of physical cells, its components, or collections, (D) the AIVC comprises virtual instruments. On the cellular scale, for example, manipulator VIs simulate how cell states change as cells divide, migrate, develop from progenitor states, or respond to perturbations through learned transitions in the URs. Decoder VIs allow for the decoding of the cell UR, e.g., to understand phenotypic properties.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

nature methods

Brief Communication

https://doi.org/10.1038/s41592-025-02772-6

Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines

Received: 11 October 2024

Accepted: 24 June 2025

Published online: 4 August 2025

Check for updates

Constantin Ahlmann-Eltze [1,2,3] ✉, Wolfgang Huber [2] & Simon Anders [1]

Recent research in deep-learning-based foundation models promises to learn representations of single-cell data that enable prediction of the effects of genetic perturbations. Here we compared five foundation models and two other deep learning models against deliberately simple baselines for predicting transcriptome changes after single or double perturbations. None outperformed the baselines, which highlights the importance of critical benchmarking in directing and evaluating method development.

The success of large language models in knowledge representation has spawned efforts to apply the foundation model concept to biology[1–3]. Several single-cell foundation models trained on transcriptomics data from millions of single cells have been published[4–6]. Two recent models—scGPT[7] and scFoundation[8]—claim to be able to predict gene expression changes caused by genetic perturbations.

In the present study, we benchmarked the performance of these models against GEARS[9] and CPA[10] and against deliberately simplistic baselines. To provide additional perspective, we also included three single-cell foundation models—scBERT[4], Geneformer[5] and UCE[6]—that were not explicitly designed for this task but can be repurposed for it by combining them with a linear decoder that maps the cell embedding to the gene expression space. In the figures, we marked their results with an asterisk.

We first assessed prediction of expression changes after double perturbations. We used data by Norman et al.[11], in which 100 individual genes and 124 pairs of genes were upregulated in K562 cells with a CRISPR activation system (Extended Data Fig. 1). The phenotypes for these 224 perturbations, plus the no-perturbation control, are logarithm-transformed RNA sequencing expression values for 19,264 genes.

We fine-tuned the models on all 100 single perturbations and on 62 of the double perturbations and assessed the prediction error on the remaining 62 double perturbations. For robustness, we ran each analysis five times using different random partitions.

For comparison, we included two simple baselines: (1) the 'no change' model that always predicts the same expression as in the control condition and (2) the 'additive' model that, for each double perturbation, predicts the sum of the individual logarithmic fold changes (LFCs). Neither uses the double perturbation data.

All models had a prediction error substantially higher than the additive baseline (Fig. 1a,b). Here, prediction error is the L_2 distance between predicted and observed expression values for the 1,000 most highly expressed genes. We also examined other summary statistics, such as the Pearson delta measure, and L_2 distances for other gene subsets: the n most highly expressed or the n most differentially expressed genes, for various n. We got the same overall result (Extended Data Fig. 2).

Next, we considered the ability of the models to predict genetic interactions. Conceptually, a genetic interaction exists if the phenotype of two (or more) simultaneous perturbations is 'surprising'. We operationalized this as double perturbation phenotypes that differed from the additive expectation more than expected under a null model with a Normal distribution (Extended Data Fig. 3 and Methods). Using the full dataset, we identified 5,035 genetic interactions (out of potentially 124,000) at a false discovery rate of 5%.

We then obtained genetic interaction predictions from each model by computing, for each of its 310,000 predictions (1,000 read-out genes and 62 held-out double perturbations across five test–training splits), the difference between predicted expression and additive expectation, and, if that difference exceeded a given threshold D, we called a predicted interaction. We then computed, for all possible choices of D, the true-positive rate (TPR) and the false discovery proportion, which resulted in the curves shown in Fig. 1c. The additive model did not compete as, by definition, it does not predict interactions.

None of the models was better than the 'no change' baseline. The same ranking of models was observed when using other metrics (Extended Data Fig. 4).

To further dissect this finding, we classified the interactions as 'buffering', 'synergistic' or 'opposite' (Fig. 1d,e and Methods). All models

[1]BioQuant, University of Heidelberg, Heidelberg, Germany. [2]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [3]Present address: UCL Cancer Institute, London, UK. ✉e-mail: constantin.ahlmann@embl.de

https://doi.org/10.1038/s41592-025-02772-6

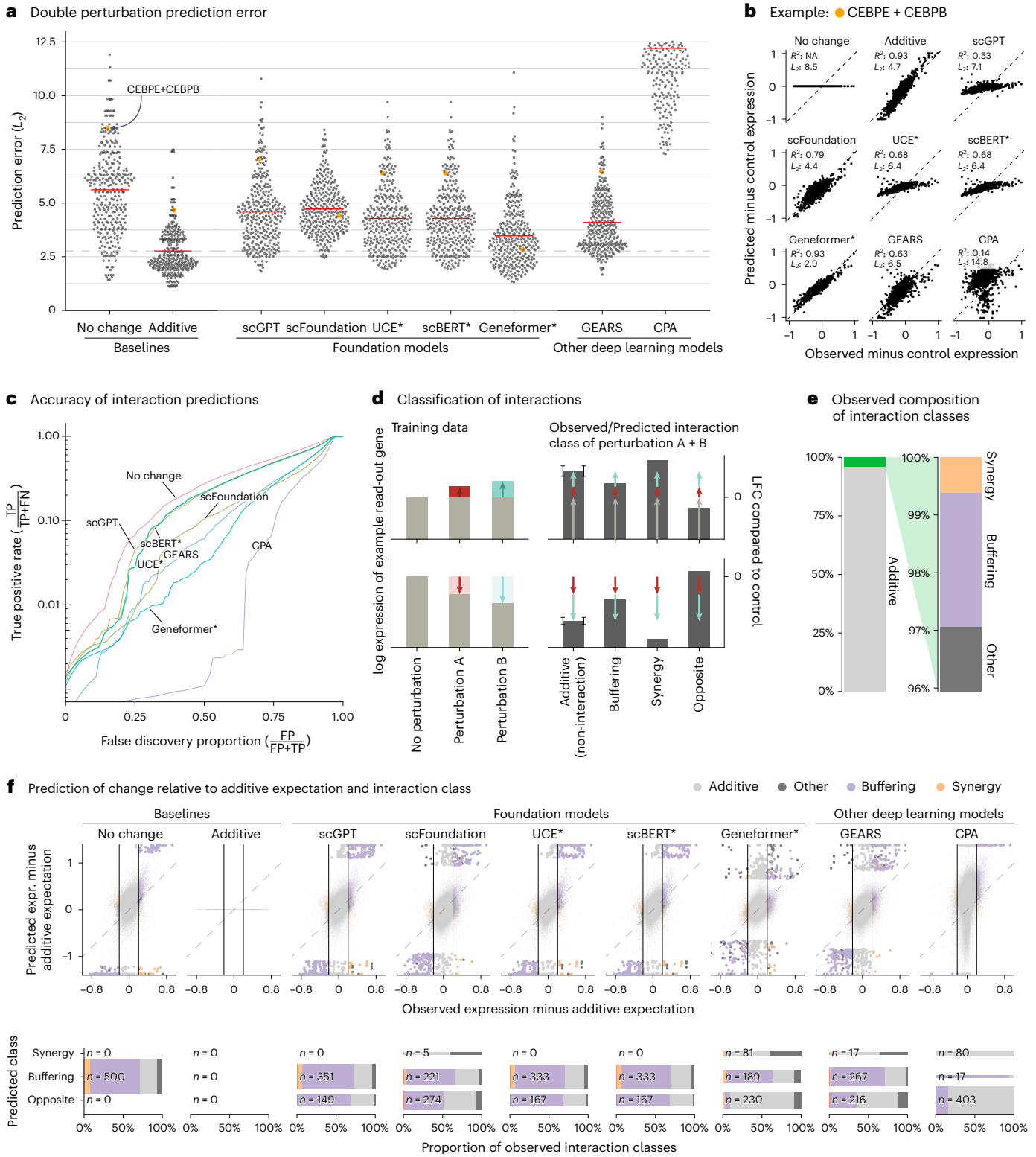


Fig. 1 | Double perturbation prediction. a, Beeswarm plot of the prediction errors for 62 double perturbations across five test–training splits. The prediction error is measured by the L_2 distance between the predicted and the observed expression profile of the $n = 1{,}000$ most highly expressed genes. The horizontal red lines show the mean per model, which, for the best-performing model, is extended by the dashed line. **b**, Scatterplots of observed versus predicted expression from one example of the 62 double perturbations. The numbers indicate error measured by the L_2 distance and the Pearson delta (R^2). **c**, TPR (recall) of the interaction predictions as a function of the false discovery proportion. FN, false negative; FP, false positive; TP, true positive. **d**, Schematic of the classification of interactions based on the difference from the additive expectation (the error bars show the additive range). **e**, Bar chart of the composition of the observed interaction classes. **f**, Top: scatterplot of observed versus predicted expression compared to the additive expectation. Each point is one of the 1,000 read-out genes under one of the 62 double perturbations across five test–training splits. The 500 predictions that deviated most from the additive expectation are depicted with bigger and more saturated points. Bottom: mosaic plots that compare the composition of highlighted predictions from the top panel stratified by the interaction class of the prediction. The width of the bars is scaled to match the number of instances. Source data for Fig. 1 are provided. expr., expression.

https://doi.org/10.1038/s41592-025-02772-6

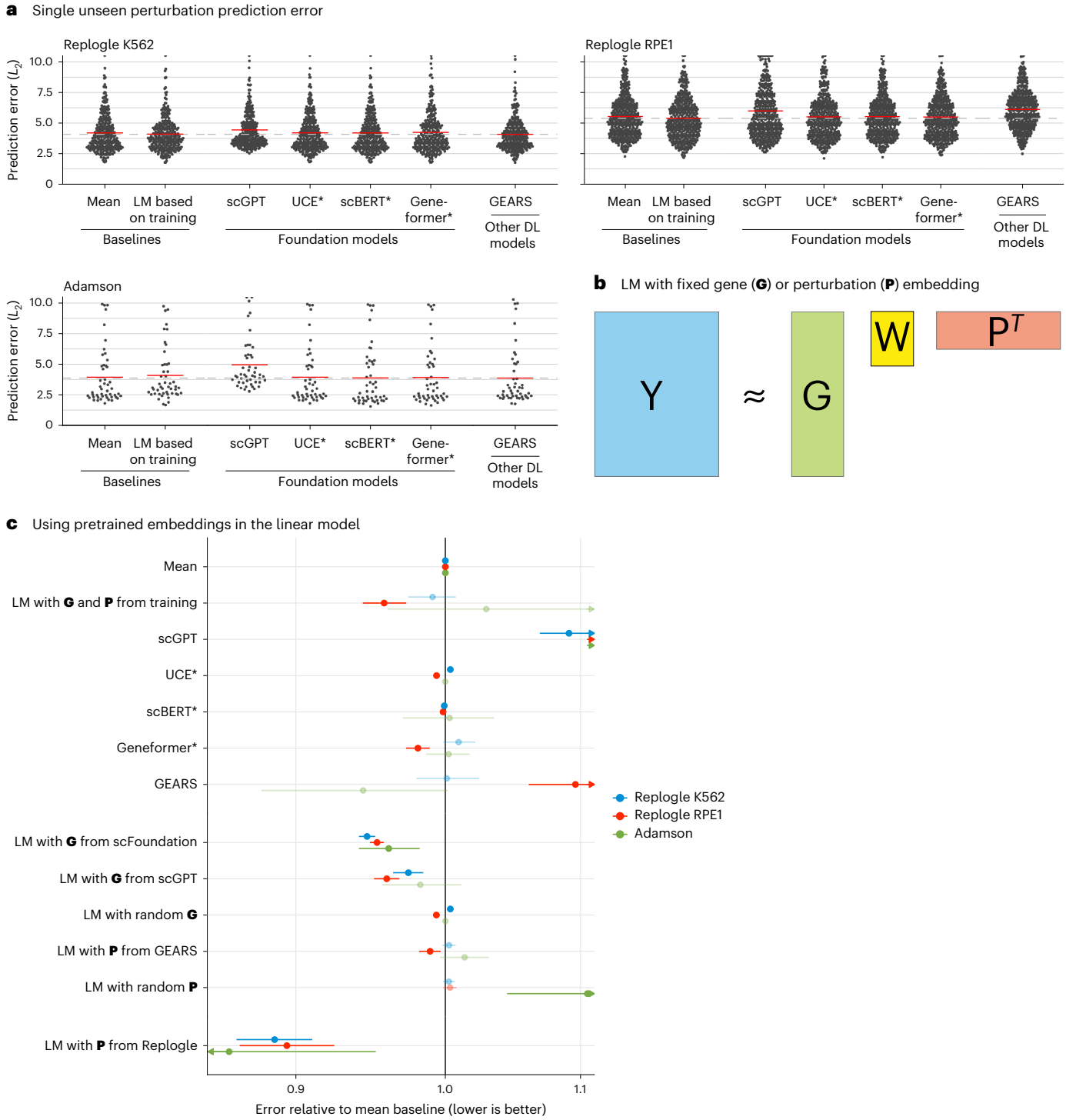


Fig. 2 | Single perturbation prediction. a, Beeswarm plot of the prediction errors for 134, 210 and 24 unseen single perturbations across two test–training splits (Methods). The prediction error is measured by the L_2 distance between the mean predicted and observed expression profile of the $n = 1{,}000$ most highly expressed genes. The horizontal red lines show the mean per model, which, for the best-performing model, is extended by the dashed line. DL, deep learning; LM, linear model. **b**, Schematic of the LM and how it can accommodate available gene (**G**) or perturbation (**P**) embeddings. **c**, Forest plot comparing the performance of all models relative to the error of the 'mean' baseline. The point ranges show the overall mean and 95% confidence interval of the bootstrapped mean ratio between each model and the baseline for 134, 210 and 24 unseen single perturbations across two test–training splits. The opacity of the point range is reduced if the confidence interval contains 0. Source data for Fig. 2 are provided.

mostly predicted buffering interactions. The 'no change' baseline cannot, by definition, find synergistic interactions, but also the deep learning models rarely predicted synergistic interactions, and it was even rarer that those predictions were correct (Fig. 1f).

To our surprise, we often found the same pair of hemoglobin genes (*HBG2* and *HBZ*) among the top predicted interactions, across models and double perturbations (Extended Data Fig. 5). Examining the data, we noted that all models except Geneformer and scFoundation

predicted LFC ≈ 0—like the 'no change' baseline—for the double perturbation of these two genes, despite their strong individual effects (Extended Data Fig. 6). More generally, we noted that, for most genes, the predictions of scGPT, UCE and scBERT did not vary across perturbations, and those of GEARS and scFoundation varied considerably less than the ground truth (Extended Data Fig. 7).

GEARS, scGPT and scFoundation also claim the ability to predict the effect of unseen perturbations. GEARS uses shared Gene Ontology[12] annotations to extrapolate from the training data, whereas the foundation models are supposed to have learned the relationships between genes during pretraining to predict unseen perturbations.

To benchmark this functionality, we used two CRISPR interference datasets by Replogle et al.[13] obtained with K562 and RPE1 cells and a dataset by Adamson et al.[14] obtained with K562 cells (Extended Data Fig. 1).

As a baseline, we devised a simple linear model. It represents each read-out gene with a K-dimensional vector and each perturbation with an L-dimensional vector. These vectors are collected in the matrices $\mathbf{G}$, with one row per read-out gene, and $\mathbf{P}$, with one row per perturbation. $\mathbf{G}$ and $\mathbf{P}$ are either obtained as dimension-reducing embeddings of the training data (Methods) or provided by an external source (see below). Then, given a data matrix $\mathbf{Y}_{\text{train}}$ of gene expression values, with one row per read-out gene and one column per perturbation (that is, per condition pseudobulk of the single-cell data), the $K \times L$ matrix $\mathbf{W}$ is found as

$$\underset{\mathbf{W}}{\operatorname{argmin}} ||\mathbf{Y}_{\text{train}} - (\mathbf{GWP}^T + \boldsymbol{b})||_2^2, \quad (1)$$

where $\boldsymbol{b}$ is the vector of row means of $\mathbf{Y}_{\text{train}}$ (Fig. 2b).

We also included an even simpler baseline, $\boldsymbol{b}$, the mean across the perturbations in the training set, following the preprints by Kernfeld et al.[15] and Csendes et al.[16] that appeared while this paper was in revision.

None of the deep learning models was able to consistently outperform the mean prediction or the linear model (Fig. 2a and Extended Data Fig. 8). We did not include scFoundation in this benchmark, as it required each dataset to exactly match the genes from its own pretraining data, and, for the Adamson and Replogle data, most of the required genes were missing. We also did not include CPA, as it is not designed to predict the effects of unseen perturbations.

Next, we asked whether we could find utility in the data representations that GEARS, scGPT and scFoundation had learned during their pretraining. We extracted a gene embedding matrix $\mathbf{G}$ from scFoundation and scGPT, respectively, and a perturbation embedding matrix $\mathbf{P}$ from GEARS. The above linear model, equipped with these embeddings, performed as well or better than scGPT and GEARS with their in-built decoders (Fig. 2c). Furthermore, the linear models with the gene embeddings from scFoundation and scGPT outperformed the 'mean' baseline, but they did not consistently outperform the linear model using $\mathbf{G}$ and $\mathbf{P}$ from the training data.

The approach that did consistently outperform all other models was a linear model with $\mathbf{P}$ pretrained on the Replogle data (using the K562 cell line data as pretraining for the Adamson and RPE1 data and the RPE1 cell line for the K562 data). The predictions were more accurate for genes that were more similar between K562 and RPE1 (Extended Data Fig. 9). Together, these results suggest that pretraining on the single-cell atlas data provided only a small benefit over random embeddings, but pretraining on perturbation data increased predictive performance.

In summary, we presented prediction tasks where current foundation models did not perform better than deliberately simplistic linear prediction models, despite significant computational expenses for fine-tuning the deep learning models (Extended Data Fig. 10). As our deliberately simple baselines are incapable of representing realistic biological complexity, yet were not outperformed by the foundation models, we conclude that the latter's goal of providing a generalizable representation of cellular states and predicting the outcome of not-yet-performed experiments is still elusive.

The publications that presented GEARS, scGPT and scFoundation included comparisons against GEARS and CPA and against a linear model. Some of these comparisons may have happened to be particularly 'easy'. For instance, CPA was never designed to predict effects of unseen perturbations and was particularly uncompetitive in the double perturbation benchmark. The linear model used in scGPT's benchmark appears to have been set up such that it reverts to predicting no change over the control condition for any unseen perturbation.

Our results are in line with previously published benchmarks that assessed the performance of foundation models for other tasks and found negligible benefits compared to simpler approaches[17–19]. Our results also concur with two previous studies showing that simple baselines outperform GEARS for predicting unseen single or double perturbations[20,21]. Since the release of our paper as a preprint, several other benchmarks[15,16,22–27] were released that also show that deep learning models struggle to outperform simple baselines. Two of these preprints[15,16] suggested an even simpler baseline than our linear model (equation (1)), namely, to always predict the overall average, and we have included this idea here.

One limitation of our benchmark is that we used only four datasets. We chose these as they were used in the publications presenting GEARS, scGPT and scFoundation. Another limitation is that all datasets are from cancer cell lines, which, for example, Theodoris et al.[5] excluded from their training data because of concerns about their high mutational burden. We also did not attempt to improve the original quality control, for example, by excluding perturbations that did not affect the expression of their own target gene and, thus, might not have worked as intended.

Deep learning is effective in many areas of single-cell omics[28,29]. However, prediction of perturbation effects still remains an open challenge, as our present work shows. We expect that increased focus on performance metrics and benchmarking will be instrumental to facilitate eventual success in applying transfer learning to perturbation data.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02772-6.

References

1. Gavriilidis, G. I., Vasileiou, V., Orfanou, A., Ishaque, N. & Psomopoulos, F. A mini-review on perturbation modelling across single-cell omic modalities. *Comput. Struct. Biotechnol. J.* **23**, 1886 (2024).
2. Szałata, A. et al. Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
3. Rood, J. E., Hupalowska, A. & Regev, A. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell* **187**, 4520–4545 (2024).
4. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
5. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
6. Rosen, Y. et al. Universal cell embeddings: a foundation model for cell biology. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.28.568918 (2023).
7. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
8. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).

9. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.* **42**, 927–935 (2024).
10. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
11. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
12. The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
13. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575 (2022).
14. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
15. Kernfeld, E., Yang, Y., Weinstock, J. S., Battle, A. & Cahan, P. A systematic comparison of computational methods for expression forecasting. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.28.551039 (2024).
16. Csendes, G., Sanz, G., Szalay, K. Z. & Szalai, B. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics* **26**, 393 (2025).
17. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* **26**, 101 (2025).
18. Boiarsky, R., Singh, N., Buendia, A., Getz, G. & Sontag, D. A deep dive into single-cell RNA sequencing foundation models. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.19.563100 (2023).
19. Liu, T., Li, K., Wang, Y., Li, H. & Zhao, H. Evaluating the utilities of foundation models in single-cell data analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.08.555192 (2024).
20. Märtens, K., Donovan-Maiye, R. & Ferkinghoff-Borg, J. Enhancing generative perturbation models with LLM-informed gene embeddings. In *Proc. Workshop on Machine Learning for Genomics Explorations* (ICLR, 2024). https://openreview.net/forum?id=eb3ndUlkt4
21. Gaudelet, T. et al. Season combinatorial intervention predictions with salt & peper. In *Proc. Workshop on Machine Learning for Genomics Explorations* (ICLR, 2024). https://openreview.net/forum?id=Wj95feICkN
22. Wenteler, A. et al. PertEval-scFM: benchmarking single-cell foundation models for perturbation effect prediction. In *Proc. 42nd International Conference on Machine Learning* (ICML, 2025).
23. Bendidi, I. et al. Benchmarking transcriptomics foundation models for perturbation analysis: one PCA still rules them all. Preprint at https://arxiv.org/abs/2410.13956 (2024).
24. Wu, Y. et al. PerturBench: benchmarking machine learning models for cellular perturbation analysis. Preprint at https://arxiv.org/abs/2408.10609 (2024).
25. Li, L. et al. A systematic comparison of single-cell perturbation response prediction models. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.23.630036 (2024).
26. Li, C. et al. Benchmarking AI models for in silico gene perturbation of cells. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.20.629581 (2025).
27. Wong, D. R., Hill, A. S. & Moccia, R. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations. *Bioinformatics* **41**, btaf317 (2025).
28. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
29. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

Methods

Data

We ran the double perturbation benchmark on the data produced by Norman et al.[11] and reprocessed by scFoundation. For the single gene perturbation benchmarks, we used the data from Adamson et al.[14] and Replogle et al.[13] as provided by GEARS (details in 'Data availability').

Software versions and parameters

We ran GEARS version 0.1.2, scGPT version 0.2.1, scFoundation (which is built on top of a GEARS version 0.0.2 fork), CPA version 0.8.8, Geneformer version 0.1.0, scBERT from commit hash 262fd4b9 with model weights provided by the authors and UCE at commit hash 8227a65c. We used each model, as much as possible, with their default parameters. All scripts that were used to predict the expression changes are available on GitHub (https://github.com/const-ae/linear_perturbation_prediction-Paper/tree/main/benchmark/src).

- GEARS and scFoundation provide a straightforward application programming interface (API) to predict the expression change after perturbation. We limited the fine-tuning time to 3 days, which meant that we trained scFoundation for five epochs.
- For scGPT, we used the same parameters and code as in their tutorial for perturbation prediction.
- For CPA, we used the code from their tutorial on how to predict combinatorial CRISPR perturbations on the Norman dataset.
- For Geneformer, we fine-tuned the provided model by predicting the perturbation labels of the training data. We then used the built-in in silico perturbation functionality to calculate the perturbed embedding.
- UCE is designed for zero-shot use, which means that it does not need to be fine-tuned. We report results from the four-layer version of UCE (as we found no performance difference between the four-layer and 33-layer versions). UCE does not provide functionality for in silico perturbation, so we calculated the post-perturbation embedding by taking the expression matrix for the unperturbed cells and overwrote the rows for the genes that we wanted to perturb with the values from the ground truth expression matrix. We, thus, tried to ensure that we tested the model under the best conditions, accepting that test data leakage could theoretically give the model an advantage over the other models.
- We fine-tuned scBERT on predicting the perturbation labels of the training data. We then used the same approach to calculate the embedding after in silico perturbation that we used for UCE.

To predict the expression changes from the embeddings of Geneformer, UCE and scBERT, we added a linear decoder to the models. We fitted a ridge regression model that predicted the gene expression of the perturbed cells from the perturbed embeddings of the training data. We then used that ridge regression to predict the gene expression of the test data from the corresponding perturbed embeddings and continued with the mean of the predicted values per perturbation.

To reduce the probability that we understate the performance for any of the models, due to wrong or suboptimal operation by ourselves, we reached out to the original authors of the benchmarked models and asked them to review our code. The authors of CPA perceived a problem with our code and submitted a fix; however, as the new code had worse performance than the original version, here we report results of the original code.

Double perturbation benchmark setup

For the double perturbation benchmark, we split the data into test and training sets. We assigned all single-gene perturbations and a randomly chosen half of the double perturbations to the training set and used the other half of the double perturbations as the test set. To reduce stochastic effects on our results, we repeated the whole procedure, including the random test–training splitting, five times.

We used two baseline models: 'no change' and 'additive'. The 'no change' model 'predicted', for each double perturbation, the expression values seen in the control condition ($\boldsymbol{y}^{\varnothing}$). The 'additive' model predicts the expression after a double perturbation of genes A and B as

$$\hat{\boldsymbol{y}}^{\text{add}} = \boldsymbol{y}^{A} + \boldsymbol{y}^{B} - \boldsymbol{y}^{\varnothing}, \tag{2}$$

where $\boldsymbol{y}^{A}$ and $\boldsymbol{y}^{B}$ are the mean observed expression vectors for the single perturbation of genes A and B, respectively.

We defined genetic interactions as follows. For each of the 124 double perturbations and the 1,000 read-out genes, we computed the difference between the observed expression value and the additive expectation. These values showed a mixture distribution composed of a large component with a single narrow peak around 0 (corresponding to a majority of non-interactions) and a smaller component consisting of two pronounced tails on either side (corresponding to interactions) (Extended Data Fig. 3). To decompose this mixture, we used Efron's empirical null approach[30], as implemented in the 'locfdr' package (version 1.1-8).

We further classified the interactions, if the two individual LFCs had the same sign, as:

- 'buffering', if the LFC was between 0 and the additive expectation
- 'synergistic', if it exceeded the additive expectation
- 'opposite', if its sign differed from that of the individual perturbations

If the individual effects were in opposite directions, 'other'. According to this classification, 2.3% of the read-out gene expression values across all double perturbation were buffering interactions; 0.6% were synergistic; and zero were in the opposite direction of the individual perturbations.

Single perturbation benchmark setup

For the single perturbation benchmark, we used the data as provided by GEARS and also used its 'simulation' test–training splitting procedure, which we repeated twice.

To predict the effects of unseen single perturbations, we used two baselines. The 'mean' model calculated the mean of the expression values in the training data. The 'linear model' is implied in equation (1). We set $\boldsymbol{b}$ to the row means of the training data ($\boldsymbol{b} = 1/N\sum_i \mathbf{Y}^{\text{train}}_{:i}$). We find $\mathbf{G}$ and $\mathbf{P}$ as follows. Perform a principal component analysis (PCA) on $\mathbf{Y}_{\text{train}}$ and use the top K principal components for $\mathbf{G}$. Then, subset this $\mathbf{G}$ to only the rows corresponding to genes that were perturbed in the training data (and, hence, appear as columns in $\mathbf{Y}^{\text{train}}$) and use the resulting matrix for $\mathbf{P}$.

Then, we find $\mathbf{W}$ using the normal equations

$$\mathbf{W} = (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^T(\mathbf{Y}_{\text{train}} - \boldsymbol{b})\mathbf{P}(\mathbf{P}^T\mathbf{P} + \lambda\mathbf{I})^{-1}, \tag{3}$$

where we use a ridge penalty of $\lambda = 0.1$ for numerical stability. Having found a $\mathbf{W}$, we can use it for prediction, $\hat{\mathbf{Y}} = \mathbf{G}\mathbf{W}\tilde{\mathbf{P}}^T + \boldsymbol{b}$, where now $\tilde{\mathbf{P}}$ is the matrix formed by the rows of $\mathbf{G}$ corresponding to genes perturbed in the test data.

For the single perturbation analysis, not all models were able to predict the expression change for all unseen perturbations. For example, the linear model with $\mathbf{G}$ and $\tilde{\mathbf{P}}$ from the training data could only predict perturbations where the target genes were also part of the read-out genes. To evaluate all models on a consistent set of perturbations, we restricted our analysis to those perturbations for which we had predictions from all models (73 perturbations for Adamson, 398 for Replogle K562 and 629 for Replogle RPE1).

We converted GEARS' Gene Ontology annotations into a perturbation embedding $\mathbf{P}$ by computing a spectral embedding[31,32] of the pathway membership matrix. We extracted the gene embedding $\mathbf{G}$ from scGPT following their tutorial on gene regulatory inference. For scFoundation, we extract $\mathbf{G}$ directly from the pretrained model weights ('pos_emb.weight'). For the

linear model with **P** from the Replogle data, we fitted a 10-dimensional PCA on the columns of the matrix with the perturbation means of the reference data. We fitted all linear models as described in the main text with $K = 10$; if **G** or **P** was provided, we simply replaced the estimate from the training data with the provided matrix before calculating **W**.

The additive model is a special case of the linear model (equation (1)) where the gene embedding is simply the single perturbation data, without any further transformation or dimension reduction ($\mathbf{G} = \mathbf{Y}^{\text{single}}$); the perturbation embedding **P** is a binary coding, where each column vector has 1s in the rows of the perturbed genes and is 0 otherwise; and W is an identity matrix and $\boldsymbol{b} = -\boldsymbol{y}^{\varnothing}$.

Evaluation metrics

We measured the prediction error using the distance $L_2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sqrt{\sum_g (\hat{y}_g - y_g)^2}$ (also called root mean squared error) between the observed expression values and predictions for the 1,000 most highly expressed genes in the control condition. We also calculated the Pearson delta correlation metric, as suggested by Cui et al.[7]: $\text{PearsonDelta}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \text{cor}(\hat{\boldsymbol{y}} - \boldsymbol{y}^{\varnothing}, \boldsymbol{y} - \boldsymbol{y}^{\varnothing})$. Unlike the L_2 distance, the Pearson delta metric does not penalize predictions that are consistently too small or too large in amplitude and, thus, prioritizes correct prediction of the direction of the expression change.

For the double perturbation data, we assess the TPR (recall) as a function of the false discovery rate. First, we find the order statistic of absolute difference of predictions and additive expectation across all test perturbations ($\boldsymbol{j} = \text{argsort}(\text{abs}(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{\text{add}}))$), where $\hat{\mathbf{Y}}$ is the matrix of the predictions for all genes and perturbations and $\hat{\mathbf{Y}}^{\text{add}}$ are the additive expectations.

The false discovery proportion (FDP) at position $l \in \{1, \cdots, N\}$ for a threshold u, which separates the interactions from the non-interactions, is

$$\text{FDP}_l = \frac{\sum_{i=1}^{l} \mathbf{1}(\text{abs}(\mathbf{Y} - \hat{\mathbf{Y}}^{\text{add}})_{\boldsymbol{j}_i} < u)}{l} \tag{4}$$

and the TPR is

$$\text{TPR}_l = \frac{\sum_{i=1}^{l} \mathbf{1}(\text{abs}(\mathbf{Y} - \hat{\mathbf{Y}}^{\text{add}})_{\boldsymbol{j}_i} \geq u)}{\sum_{i=1}^{N} \mathbf{1}(\text{abs}(\mathbf{Y} - \hat{\mathbf{Y}}^{\text{add}})_i \geq u)}, \tag{5}$$

where **Y** is the matrix of observed value and N is the product of the number of genes and perturbations. The indicator function $\boldsymbol{1}(\cdot)$ counts how often the observed values **Y** deviate enough from the additive expectation so that the observations are considered an interaction. The order statistic $\boldsymbol{j}$ ensures that we consider the gene–perturbation pairs first, where the model prediction deviates most from the additive expectation.

Lastly, we find the order statistic of the FDPs ($\boldsymbol{s} = \text{argsort}(\text{FDP})$) and plot the tuples $1, \cdots, N$

$$(\text{FDP}_{\boldsymbol{s}_i}, \max_i^{1\cdots l}(\text{TPR}_{\boldsymbol{s}_i})). \tag{6}$$

An advantage of considering here the false discovery versus true-positive curve, compared with the precision-recall or the receiver operator curve, is that it provides a direct assessment of which fraction of interactions a model identifies for a fixed fraction of false positives.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this paper are publicly available: the Norman et al.[11] was downloaded via scFoundation (https://figshare.com/ndownloader/files/44477939); the Adamson et al.[14] was downloaded via GEARS (https://dataverse.harvard.edu/api/access/datafile/6154417); the Replogle et al.[13] K562 was downloaded via GEARS (https://dataverse.harvard.edu/api/access/datafile/7458695); and the Replogle et al.[13] RPE1 was also downloaded via GEARS (https://dataverse.harvard.edu/api/access/datafile/7458694). Source data for Figs. 1 and 2 and Extended Data Figs. 1–3, 5, 6 and 8–10 are provided.

Code availability

The code to reproduce the analyses presented here and details about the software package versions are available at github.com/const-ae/linear_perturbation_prediction-Paper, which we also archived on Zenodo[33]. The Zenodo repository also contains the results of the intermediate calculations needed to reproduce all figures.

References

30. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Vol. 1 (Cambridge Univ. Press, 2012).
31. Sussman, D. L., Tang, M., Fishkind, D. E. & Priebe, C. E. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Am. Stat. Assoc.* **107**, 1119–1128 (2012).
32. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **1695**, 1–9 (2006).
33. Ahlmann-Eltze, C. Code and data of this benchmark. *Zenodo* https://doi.org/10.5281/zenodo.14832393 (2025).
34. Ahlmann-Eltze, C. & Huber, W. Analysis of multi-condition single-cell data with latent embedding multivariate regression. *Nat. Genet.* **57**, 659–667 (2025).

Acknowledgements

We thank an anonymous reviewer of one of our previous papers[34] for the suggestion to compare foundation models against simple linear models, which eventually led to this work. This work has been supported by the European Research Council (Synergy Grant DECODE under grant agreement number 810296) and by the Klaus Tschira Foundation (grant number 00.022.2019).

Author contributions

C.A.E., W.H. and S.A. conceived the study and wrote the paper. C.A.E. performed the computations, with feedback from W.H. and S.A.

Funding

Open access funding provided by European Molecular Biology Laboratory (EMBL).

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02772-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02772-6.

Correspondence and requests for materials should be addressed to Constantin Ahlmann-Eltze.

Peer review information *Nature Methods* thanks Luca Pinello and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

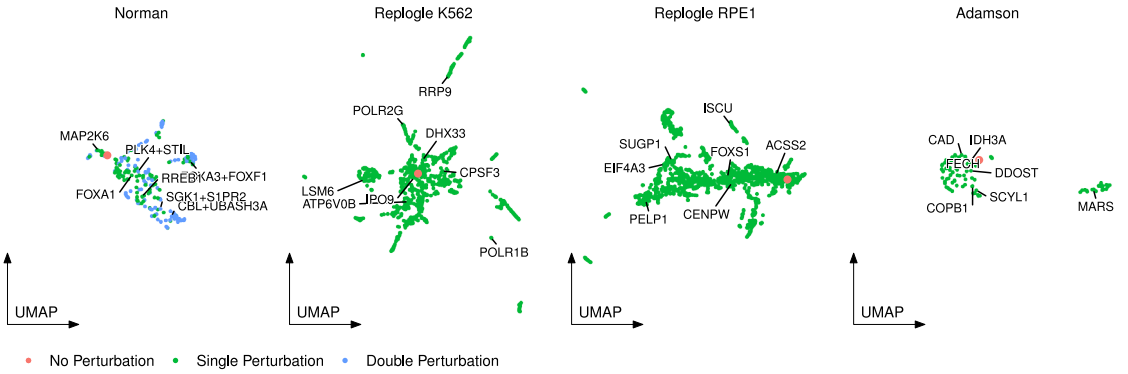
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-025-02772-6

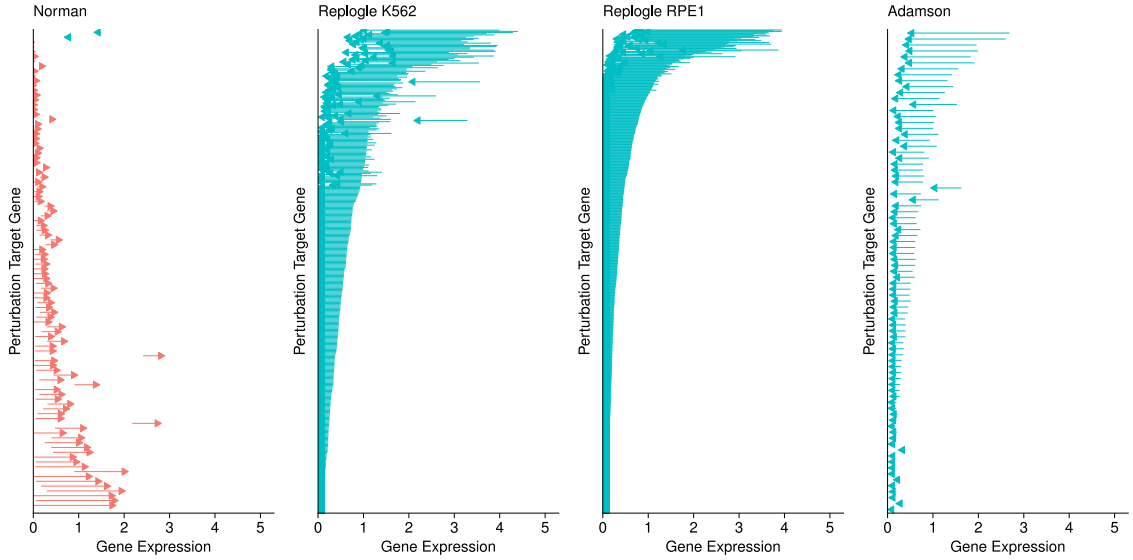
(A) Table: Dataset overview

Dataset (cell line)	Size (Genes × Cells)	Perturbations
Norman (K562)	19 264 × 84 143	124 double and 100 single perturbations
Replogle (K562)	5 000 × 162 264	1 087 single perturbations
Replogle (RPE1)	5 000 × 161 423	1 534 single perturbations
Adamson (K562)	5 060 × 65 899	81 single perturbations

(B) UMAP of the datasets

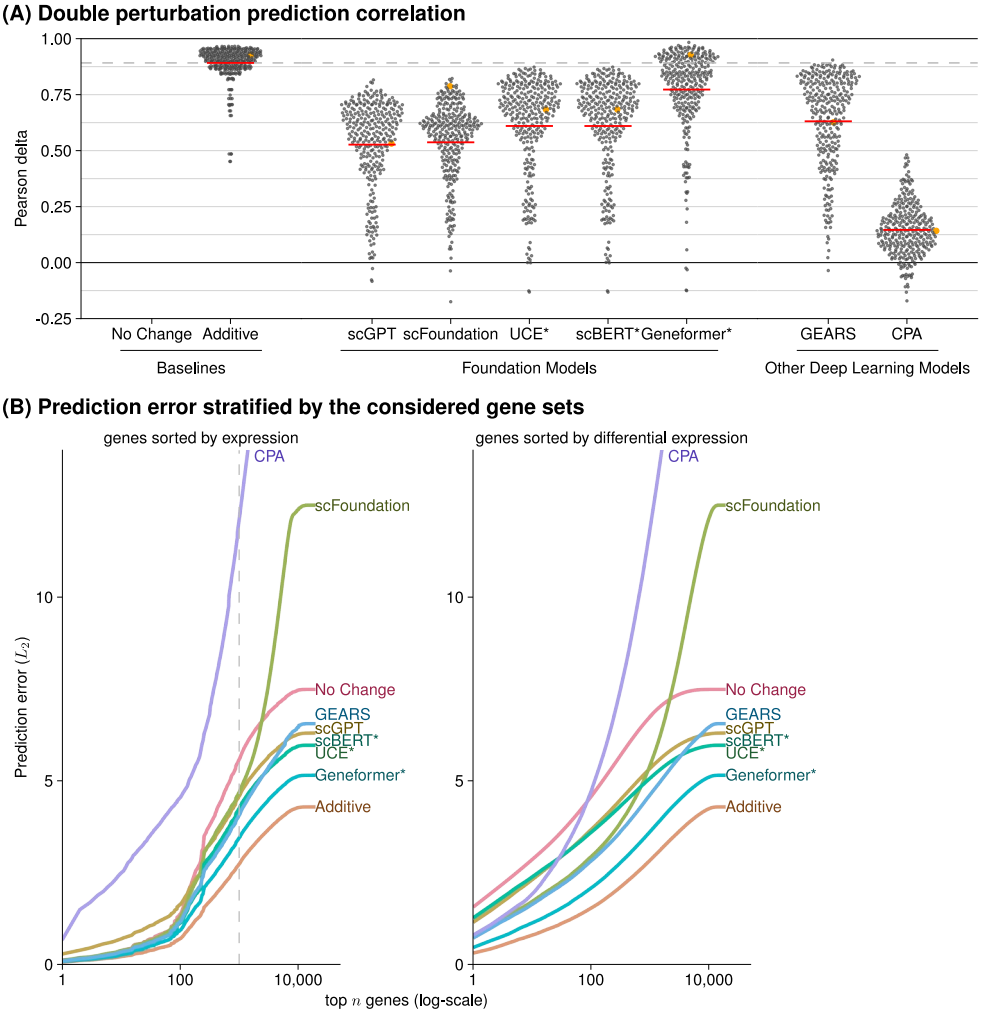


(C) Change of perturbation target gene expression



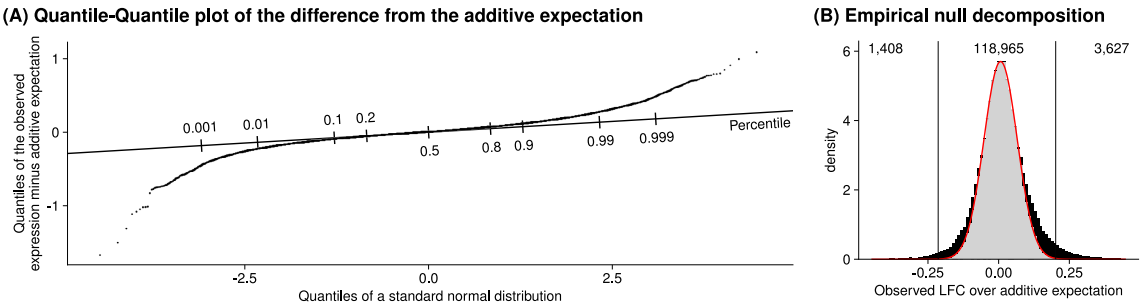
Extended Data Fig. 1 | Dataset Overview. (**a**) Table with the size of the data and the number of perturbations. (**b**) UMAP on the perturbations per dataset (aggregated to the mean per perturbation). The position of the control condition without perturbation is shown in red, and a random selection of perturbations is labeled. (**c**) Change in the expression of the target gene of each perturbation. The base of the arrow indicates the expression without perturbation, and the tip indicates the expression after perturbation. For genes targeted multiple times in the Norman dataset, we show the average expression after perturbation.

https://doi.org/10.1038/s41592-025-02772-6



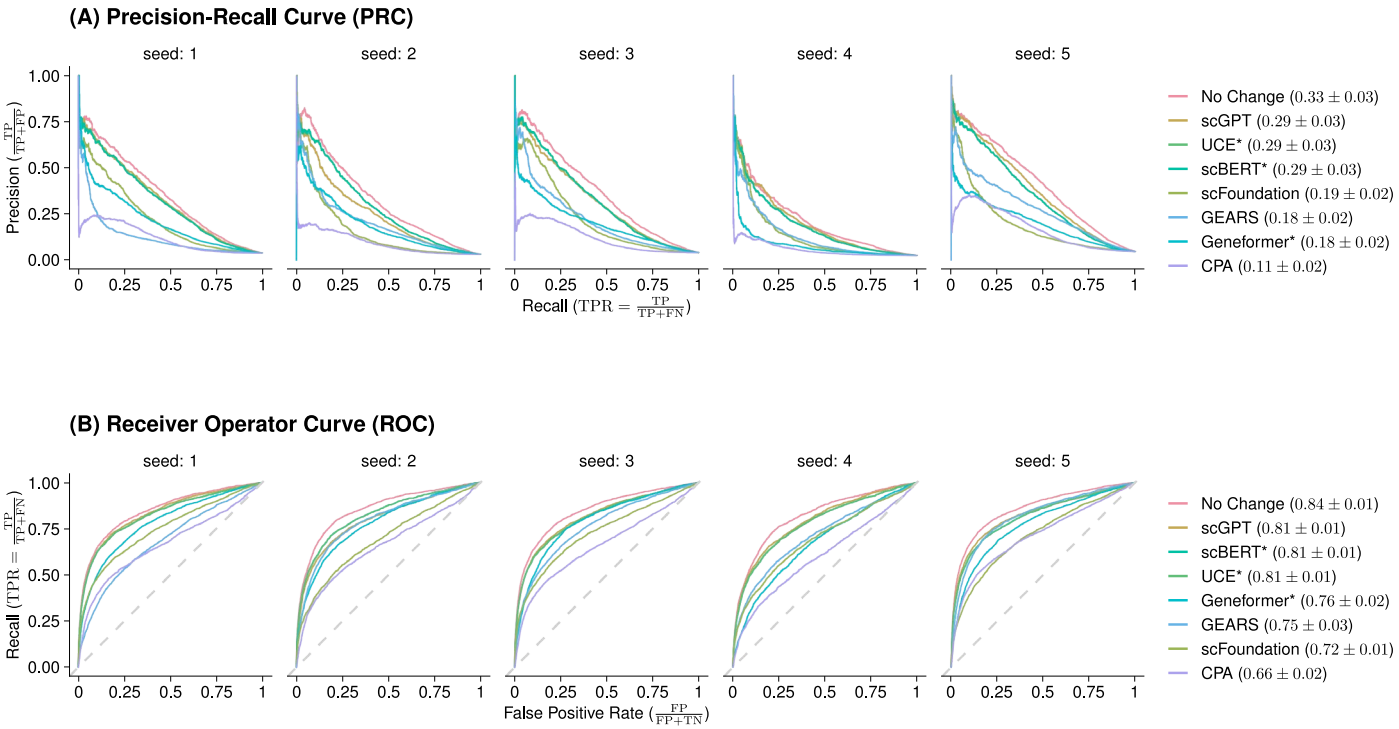
Extended Data Fig. 2 | Alternative measures of the double perturbation prediction performance. (**a**) The Pearson delta measure calculates the correlation of the prediction and observations after subtracting the expression in the control condition. The correlation for the *no change* predictions could not be calculated because they were all zero. The horizontal red lines show the mean per model, and the dashed line indicates the correlation of the best-performing model. (**b**) Prediction error as a function of n, the number of read-out genes. Left: genes ranked by expression in the control condition, right: by differential expression between observed value and expression in the control condition. Note that sorting by differential expression is only possible if access to the ground truth is available and can thus not be applied in real-world use cases. The dashed line at $n = 1000$ is the choice in Panel **a** and elsewhere in this work.

https://doi.org/10.1038/s41592-025-02772-6



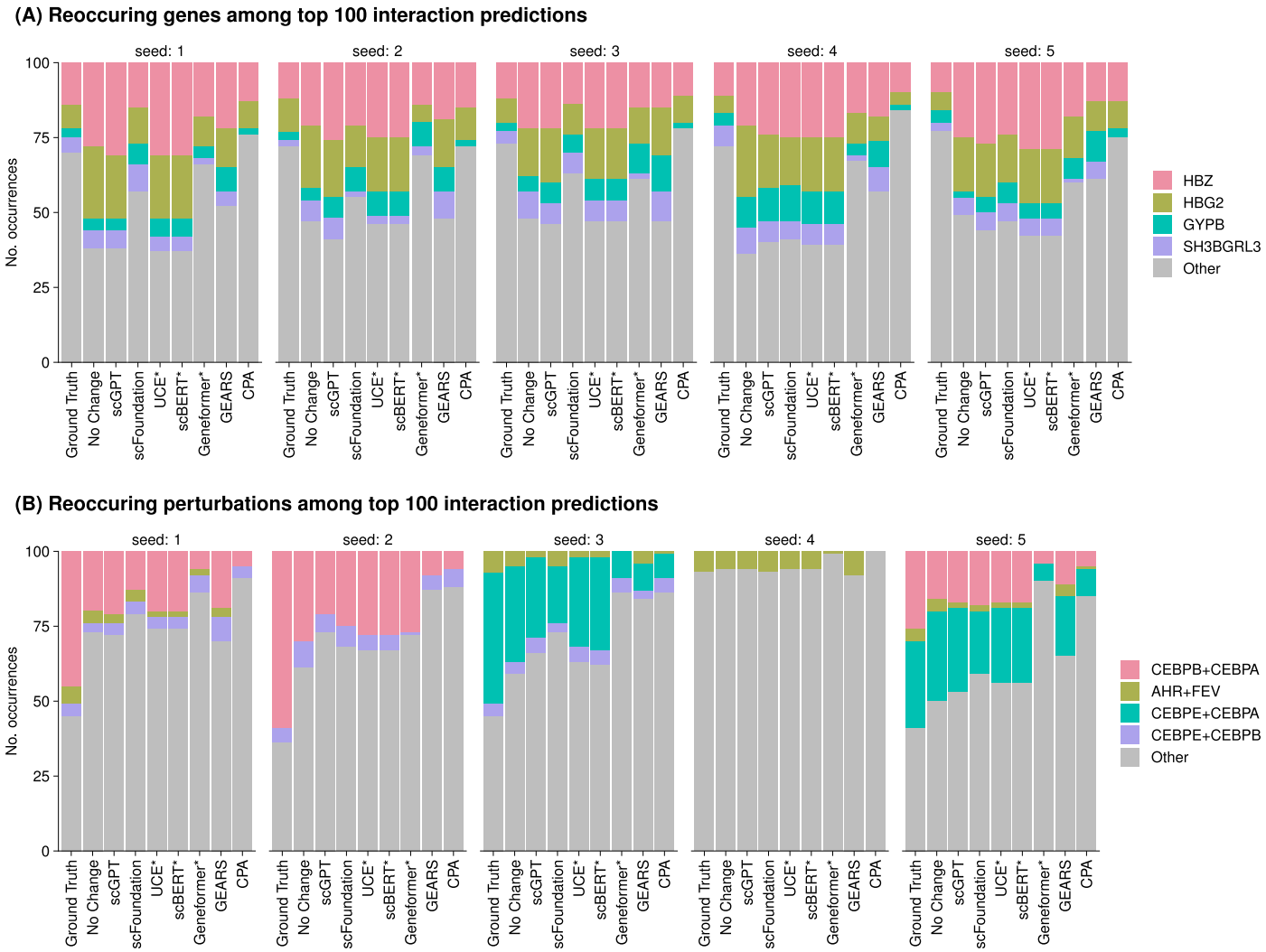
Extended Data Fig. 3 | Distribution of the observed difference from the additive model. (**a**) Quantile-quantile plot comparing the distribution of the differences between observed expression values and the additive expectation against a standard normal distribution. The slope of the line is the standard deviation of the null model. (**b**) Histogram of the differences with a red curve overlayed that shows the null distribution fitted using *locfdr*. Values under the curve are grey, and the black bars show the observations that exceed what we would expect under the null model. The vertical bar shows the upper and lower thresholds for which the observations have a false discovery rate of less than 5% (that is, the grey fraction of the bars outside the vertical lines is 5%). The numbers at the top count the observations per group.

https://doi.org/10.1038/s41592-025-02772-6



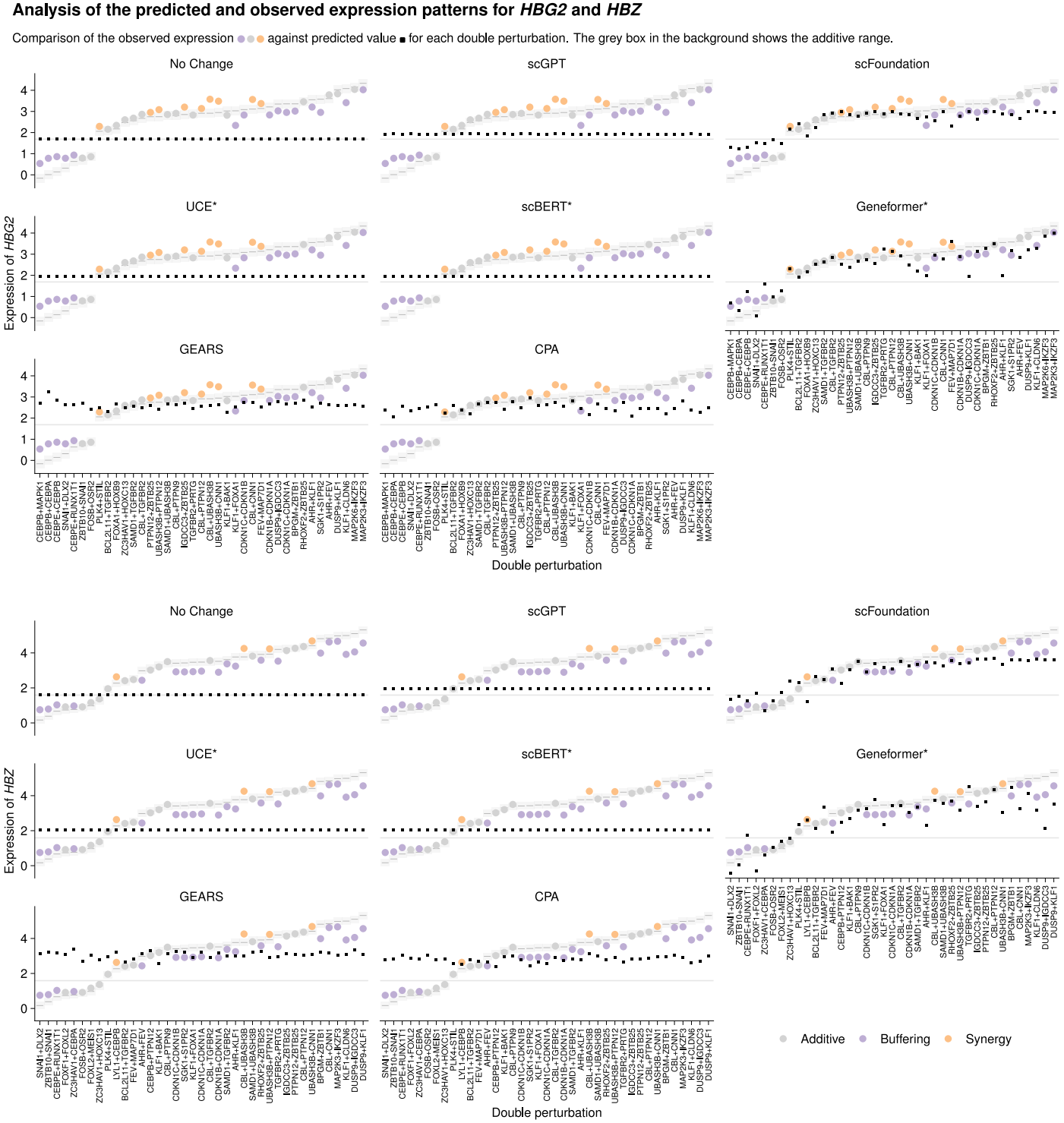
Extended Data Fig. 4 | Alternative measures how well each model detects genetic interactions. (**a**) Precision-recall and (**b**) receiver operator curve for all models distinguishing interactions from additive combinations. The numbers in parenthesis are the area under the curve (AUC) with the standard error across five test-training splits. *TP*: true positive, *FP*: false positive, *FN*: false negative, *TN*: true negative.

https://doi.org/10.1038/s41592-025-02772-6



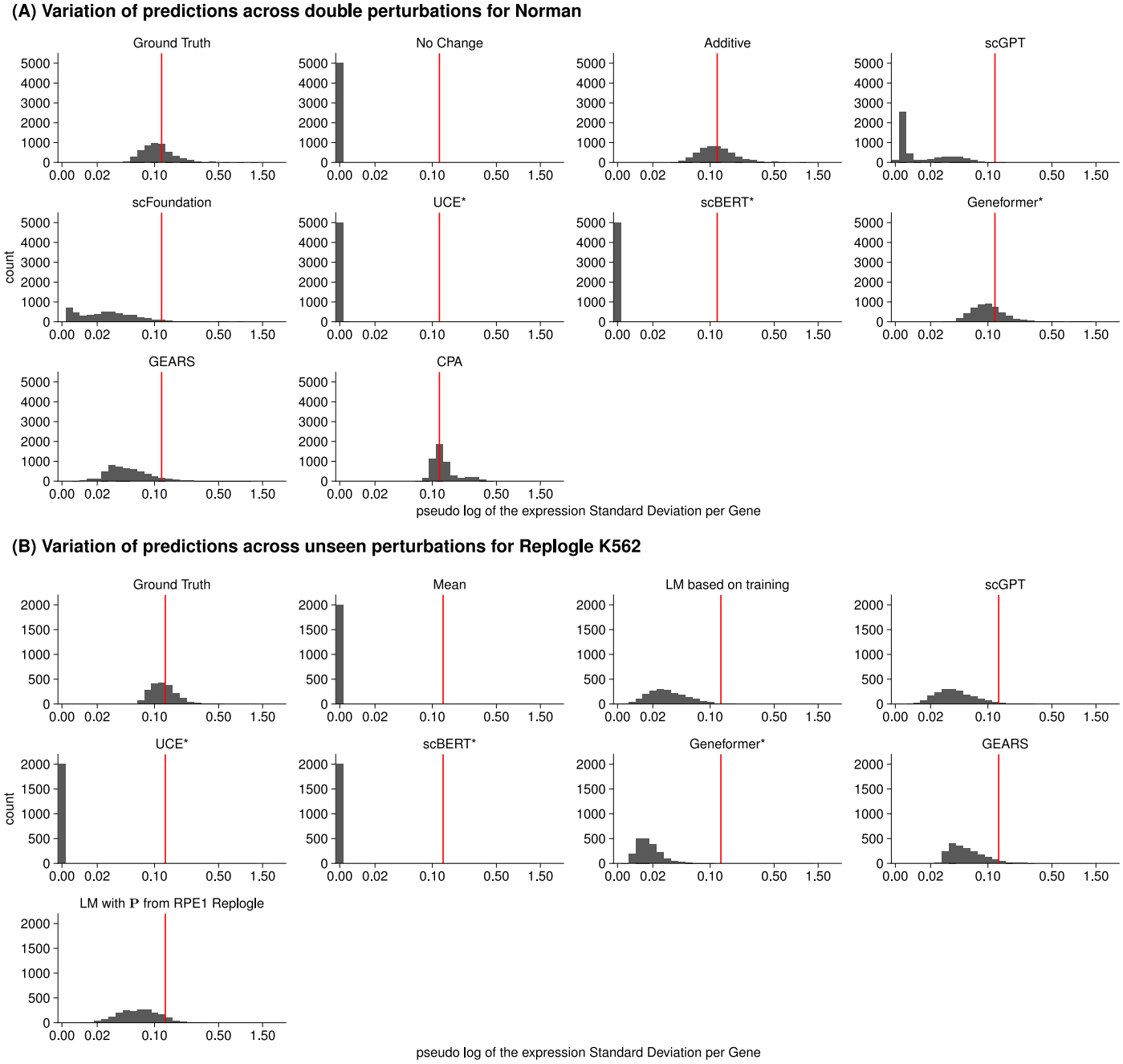
Extended Data Fig. 5 | Reoccurrence of genes and perturbations for top predictions. (**a**) Reoccurrence of genes and (**b**) perturbations among the 100 predictions that differed most from the additive expectation. The data is facetted by the test-training split. The ground truth column shows the genes and perturbations sorted by observed difference from the additive expectation. The highlighted genes and perturbations are the four most reoccurring ones.

https://doi.org/10.1038/s41592-025-02772-6



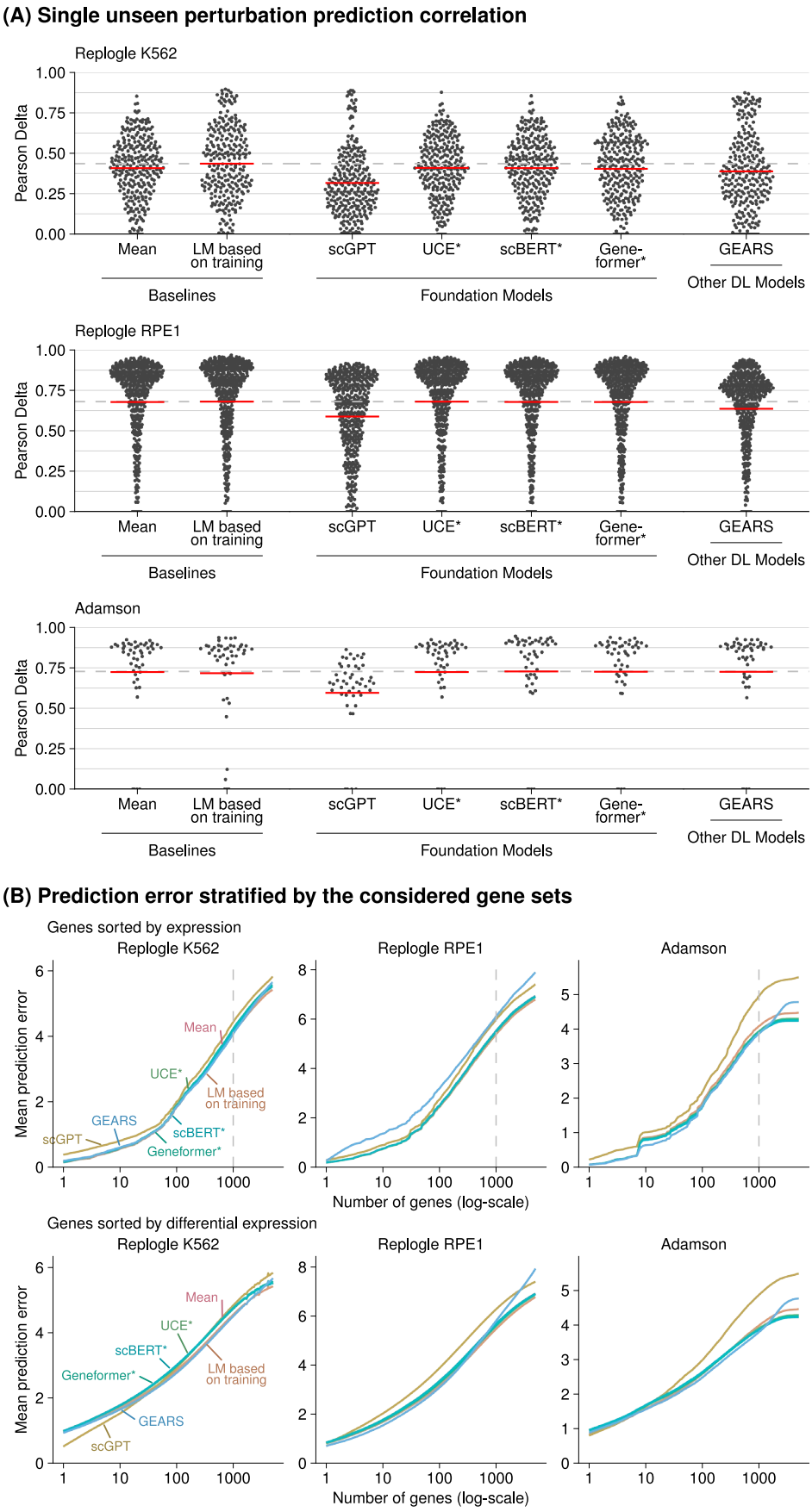
Extended Data Fig. 6 | Comparison of predicted and observed expression for *HBG2* and *HBZ*. Comparison of the predicted expression (black squares), the observed expression values (points colored by interaction type), and the range of values that are considered additive (grey boxes) for all test perturbations with seed = 1. The grey horizontal line shows the expression of *HBG2* and *HBZ* without perturbation.

https://doi.org/10.1038/s41592-025-02772-6



Extended Data Fig. 7 | Variation of the predicted and observed expression values. Histogram of the standard deviation per gene for the predicted and observed expression values across perturbations facetted by the model. The red vertical bar indicates the mean of the standard deviations for the ground truth for (**a**) the Norman dataset and (**b**) the Replogle K562 dataset. The data reflects the variation for the 1 000 most highly expressed genes and is aggregated across five test-training splits. *LM*: linear model.

https://doi.org/10.1038/s41592-025-02772-6

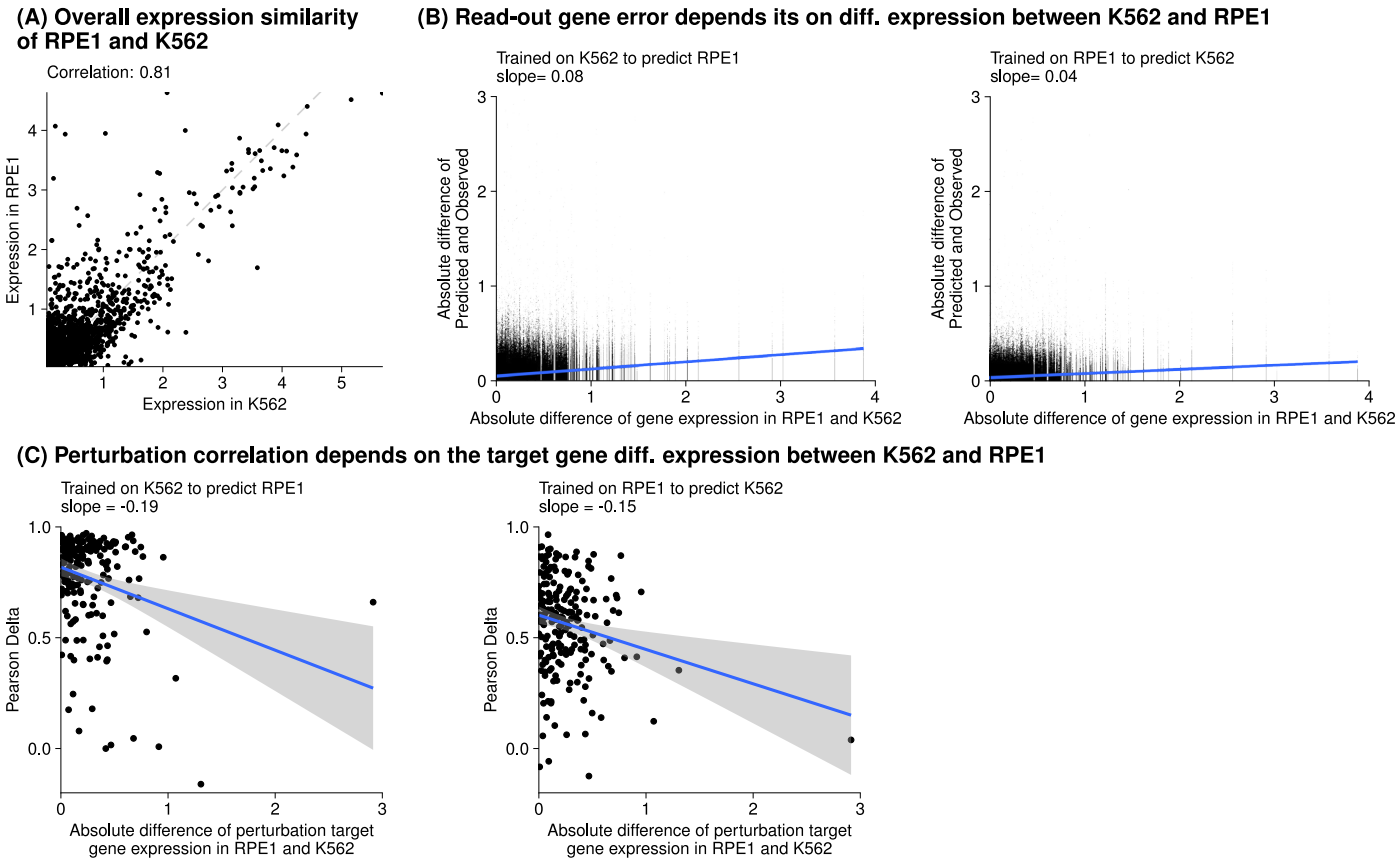


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41592-025-02772-6

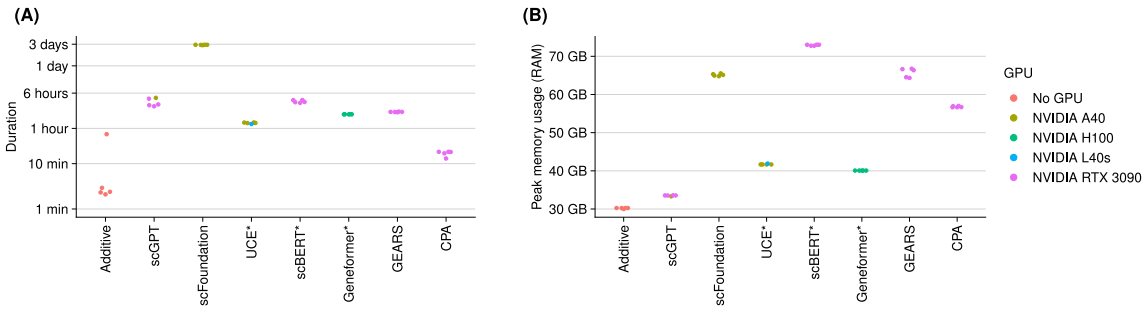
Extended Data Fig. 8 | Alternative measures of the single perturbation prediction performance. (**a**) The Pearson delta measure calculates the correlation of the prediction and observations after subtracting the expression in the control condition. The horizontal red lines show the mean per model and the dashed line indicates the correlation of the best-performing model. (**b**) Prediction error as a function of *n*, the number of read-out genes. Top: genes ranked by expression in the control condition. Bottom: by differential expression between observed value and expression in the control condition. Note that sorting by differential expression is only possible if access to the ground truth is available and can thus not be applied in real-world use cases. The dashed line at $n = 1000$ is the choice in Panel **a** and elsewhere in this work. *LM*: linear model, *DL*: deep learning.

https://doi.org/10.1038/s41592-025-02772-6



Extended Data Fig. 9 | Analysis how differential expression between K562 and RPE1 effects prediction accuracy of transfer learning. (**a**) Scatter plot of the mean gene expression for shared genes between RPE1 and K562 without perturbation. The dashed line indicates the diagonal. (**b**) Scatter plot of the absolute prediction error per read-out gene against the differential expression of that gene between RPE1 and K562. Each point is one read-out gene from one of the 122 double perturbations from five test-training splits. The blue line shows the linear fit with a slope indicated in the subtitle. (**c**) Scatter plot of the Pearson delta score per perturbation for the RPE1 dataset against the differential expression of the perturbation target gene between RPE1 and K562. The blue line shows the linear fit with a slope indicated in the subtitle, and the shaded area indicates the standard error of the fit.

https://doi.org/10.1038/s41592-025-02772-6



Extended Data Fig. 10 | Computational resource requirements. The resource usage was measured for the Norman dataset, which had 19 624 genes and 81 143 cells grouped into 225 conditions. Each point is one of the five test-training splits. (**a**) Elapsed time on a log scale to fine-tune and predict the double perturbations. (**b**) Peak memory usage for each model was measured using GNU time. The points are colored by the respective GPU model that was used.

nature portfolio

Corresponding author(s): *Double-anonymous peer review submissions: write DAPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We did not use any software for data collection.
Data analysis	We used R version 4.4.1 and Python version 3.10.4 to run GEARS version 0.1.2, scGPT version 0.2.1, scFoundation (build around GEARS version 0.0.2), CPA version 0.8.8, Geneformer version 0.1.0, scBERT from commit hash 262fd4b9 with model weights provided by the authors, and UCE at commit hash 8227a65c. We used igraph version 2.1.4 and locfdr 1.1-8.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Dataset	Availability (with link)
Norman	Downloaded from scFoundation via Figshare (https://figshare.com/articles/dataset/scFoundation_Large_Scale_Foundation_Model_on_Single-

cell_Transcriptomics_-_processed_datasets/24049200?file=44477939)
Adamson Downloaded via GEARS from Harvard dataverse (https://dataverse.harvard.edu/api/access/datafile/6154417)
Replogle K562 Downloaded via GEARS from Harvard dataverse (https://dataverse.harvard.edu/api/access/datafile/7458695)
Replogle RPE1 Downloaded via GEARS from Harvard dataverse (https://dataverse.harvard.edu/api/access/datafile/7458694)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We chose publicy available datasets whose experimental designs, incl. sample size, were appropriate as a test case for the proposed method / underlying scientific question.
Data exclusions	We did not exclude any data.
Replication	All material to replicate our results are available at https://github.com/const-ae/linear_perturbation_prediction-Paper/ and https://doi.org/10.5281/zenodo.14833202. However, we did not attempt to independently replicate our results.
Randomization	Not applicable. We present a complete combinatorial matrix of benchmarks of multiple computational methods each applied to multiple datasets and thus could observe each software in all conditions and did not need to randomize software to condition assignment.
Blinding	The analysts were not blinded while evaluating the benchmark. During the implementation of the benchmark, it is not possible to be blinded, as each method requires custom code. And while reporting the results blinding was not necessary as we directly provide the results for metrics that have previously already been used to measure the accuracy of perturbation effect prediction. To ensure that all methods were fairly evaluated, we approached the authors of each method and asked them to scrutinize our implementation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

B Further papers

For each chapter, a longer list of papers grouped by theme: classic papers, reviews, methods and recent developments. They are not required reading, but they are the papers behind the text of the chapter and a good place to go deeper. Each paper is given with its title, a DOI link and, where available, its abstract.

B.1 Sequence searching

Further papers for Chapter 1.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.1.1 Reviews — start here

- Pearson (2013) — an introduction to similarity searching, homology inference and statistics; the best single starting point.

  An Introduction to Sequence Similarity ("Homology") Searching. Pearson (2013). *Current Protocols in Bioinformatics*. doi:10.1002/0471250953.bi0301s42

  > Sequence similarity searching, typically with BLAST, is the most widely used and most reliable strategy for characterizing newly determined sequences. Sequence similarity searches can identify "homologous" proteins or genes by detecting excess similarity- statistically significant similarity that reflects common ancestry. This unit provides an overview of the inference of homology from significant similarity, and introduces other units in this chapter that provide more details on effective strategies for identifying homologs.

- Eddy (2004a) — "What is a hidden Markov model?" (two pages).

  What is a hidden Markov model?. Eddy (2004). *Nature Biotechnology*. doi:10.1038/nbt1004-1315

- Eddy (1998) — profile hidden Markov models.

  Profile hidden Markov models.. Eddy (1998). *Bioinformatics*. doi:10.1093/bioinformatics/14.9.755

  > The recent literature on profile hidden Markov model (profile HMM) methods and software is reviewed. Profile HMMs turn a multiple sequence alignment into a position-specific scoring system suitable for searching databases for remotely homologous sequences. Profile HMM analyses complement standard pairwise comparison methods for large-scale sequence analysis. Several software implementations and two large libraries of profile HMMs of common protein domains are available. HMM methods performed comparably to threading methods in the CASP2 structure prediction exercise.

- Lambert et al. (2003) — review of common sequence alignment methods.

  Review of common sequence alignment methods: clues to enhance reliability. Lambert et al. (2003). *Current Genomics*.

- Bepler and Berger (2021) — learning the protein language: evolution, structure and function.

  Learning the protein language: Evolution, structure, and function. Bepler & Berger (2021). *Cell Systems*. doi:10.1016/j.cels.2021.05.017

  > Language models have recently emerged as a powerful machine-learning approach for distilling information from massive protein sequence databases. From readily available sequence data alone, these models discover evolutionary, structural, and functional organization across protein space. Using language models, we can encode amino-acid sequences into distributed vector representations that capture their structural and functional properties, as well as evaluate the evolutionary fitness of sequence variants. We discuss recent advances in protein language modeling and their applications to downstream protein property prediction problems. We then consider how these models can be enriched with prior biological knowledge and introduce an approach for encoding protein structural knowledge into the learned representations. The knowledge distilled by these models allows us to improve downstream function prediction through transfer learning. Deep protein language models are revolutionizing protein biology. They suggest new ways to approach protein and therapeutic design. However, further developments are needed to encode strong biological priors into protein language models and to increase their accessibility to the broader community.

B.1.2 Homology

- Reeck et al. (1987) — "homology" in proteins and nucleic acids: a terminology muddle and a way out of it.

  "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. Reeck et al. (1987). *Cell*. doi:10.1016/0092-8674(87)90322-9

- Doolittle (1981) — similar amino acid sequences: chance or common ancestry?

  Similar Amino Acid Sequences: Chance or Common Ancestry?. Doolittle (1981). *Science*. doi:10.1126/science.7280687

  > The systemic comparison of every newly determined amino acid sequence with all other known sequences may allow a complete reconstruction of the evolutionary events leading to contemporary proteins. But sometimes the surviving similarities are so vague that even computer-based sequence comparisons procedures are unable to validate relationships. In other cases similar sequences may appear in totally alien proteins as a result of mere chance or, occasionally, by the convergent evolution of sequences with special properties.

- Rost (1999) — the twilight zone of protein sequence alignments.

  Twilight zone of protein sequence alignments. Rost (1999). *Protein Engineering, Design and Selection*. doi:10.1093/protein/12.2.85

  > Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20-35% sequence identity. Here, more than a million sequence alignments were analysed between protein pairs of known structures to re-define a line distinguishing between true and false positives for low levels of similarity. Four results stood out. (i) The transition from the safe zone of sequence alignment into the twilight zone is described by an explosion of false negatives. More than 95% of all pairs detected in the twilight zone had different structures. More precisely, above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were. (ii) Whether or not sequence homology implied structural identity depended crucially on the alignment length. For example, if 10 residues were similar in an alignment of length 16 (>60%), structural similarity could not be inferred. (iii) The 'more similar than identical' rule (discarding all pairs for which percentage similarity was lower than percentage identity) reduced false positives significantly. (iv) Using intermediate sequences for finding links between more distant families was almost as successful: pairs were predicted to be homologous when the respective sequence families had proteins in common. All findings are applicable to automatic database searches.

- Fitch (1970) — distinguishing homologous from analogous proteins (orthology and paralogy defined).

  Distinguishing Homologous from Analogous Proteins. Fitch (1970). *Systematic Zoology*. doi:10.2307/2412448

B.1.3 Pairwise alignment and substitution matrices

- Bellman (1954) — the theory of dynamic programming.

  The theory of dynamic programming. Bellman (1954). *Bulletin of the American Mathematical Society*. doi:10.1090/s0002-9904-1954-09848-8

- Needleman and Wunsch (1970) — global alignment by dynamic programming.

  A general method applicable to the search for similarities in the amino acid sequence of two proteins. Needleman & Wunsch (1970). *Journal of Molecular Biology*. doi:10.1016/0022-2836(70)90057-4

- Gotoh (1982) — affine gap penalties in quadratic time.

  An improved algorithm for matching biological sequences. Gotoh (1982). *Journal of Molecular Biology*. doi:10.1016/0022-2836(82)90398-9

- Myers and Miller (1988) — optimal alignments in linear space.

  Optimal alignments in linear space. Myers & Miller (1988). *Bioinformatics*. doi:10.1093/bioinformatics/4.1.11

  > Space, not time, is often the limiting factor when computing optimal sequence alignments, and a number of recent papers in the biology literature have proposed space-saving strategies. However, a 1975 computer science paper by Hirschberg presented a method that is superior to the new proposals, both in theory and in practice. The goal of this paper is to give Hirschberg's idea the visibility it deserves by developing a linear-space version of Gotoh's algorithm, which accommodates affine gap penalties. A portable C-software package implementing this algorithm is available on the BIONET free of charge.

- Dayhoff et al. (1978) — PAM matrices.

  A model of evolutionary change in proteins. Dayhoff et al. (1978). *Atlas of Protein Sequence and Structure*.

- Henikoff and Henikoff (1992) — BLOSUM matrices.

  Amino acid substitution matrices from protein blocks.. Henikoff & Henikoff (1992). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.89.22.10915

  > Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.

- Altschul (1991) — substitution matrices from an information-theoretic perspective.

  Amino acid substitution matrices from an information theoretic perspective. Altschul (1991). *Journal of Molecular Biology*. doi:10.1016/0022-2836(91)90193-a

> Protein sequence alignments have become an important tool for molecular biologists. Local alignments are frequently constructed with the aid of a "substitution score matrix" that specifies a score for aligning each pair of amino acid residues. Over the years, many different substitution matrices have been proposed, based on a wide variety of rationales. Statistical results, however, demonstrate that any such matrix is implicitly a "log-odds" matrix, with a specific target distribution for aligned pairs of amino acid residues. In the light of information theory, it is possible to express the scores of a substitution matrix in bits and to see that different matrices are better adapted to different purposes. The most widely used matrix for protein sequence comparison has been the PAM-250 matrix. It is argued that for database searches the PAM-120 matrix generally is more appropriate, while for comparing two specific proteins with suspected homology the PAM-200 matrix is indicated. Examples discussed include the lipocalins, human alpha 1 B-glycoprotein, the cystic fibrosis transmembrane conductance regulator and the globins.

- Eddy (2004b) — where did the BLOSUM62 matrix come from?

  Where did the BLOSUM62 alignment score matrix come from?. Eddy (2004). *Nature Biotechnology*. doi:10.1038/nbt0804-1035

  > Many sequence alignment programs use the BLOSUM62 score matrix to score pairs of aligned residues. Where did BLOSUM62 come from?

B.1.4 Heuristic search and statistics

- Pearson and Lipman (1988) — FASTA.

  Improved tools for biological sequence comparison.. Pearson & Lipman (1988). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.85.8.2444

  > We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive derivative of the FASTP program, which can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The RDF2 program can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be displayed as a "graphic matrix" plot or as individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices.

- Altschul et al. (1990) — the original BLAST.

  Basic local alignment search tool. Altschul et al. (1990). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(05)80360-2

  > A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

- Karlin and Altschul (1990) — statistical significance of local alignment scores.

  Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.. Karlin & Altschul (1990). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.87.6.2264

> An unusual pattern in a nucleic acid or protein sequence or a region of strong similarity shared by two or more sequences may have biological significance. It is therefore desirable to know whether such a pattern can have arisen simply by chance. To identify interesting sequence patterns, appropriate scoring values can be assigned to the individual residues of a single sequence or to sets of residues when several sequences are compared. For single sequences, such scores can reflect biophysical properties such as charge, volume, hydrophobicity, or secondary structure potential; for multiple sequences, they can reflect nucleotide or amino acid similarity measured in a wide variety of ways. Using an appropriate random model, we present a theory that provides precise numerical formulas for assessing the statistical significance of any region with high aggregate score. A second class of results describes the composition of high-scoring segments. In certain contexts, these permit the choice of scoring systems which are "optimal" for distinguishing biologically relevant patterns. Examples are given of applications of the theory to a variety of protein sequences, highlighting segments with unusual biological features. These include distinctive charge regions in transcription factors and protooncogene products, pronounced hydrophobic segments in various receptor and transport proteins, and statistically significant subalignments involving the recently characterized cystic fibrosis gene.

- Kent (2002) — BLAT, for fast genome-scale alignment.

  BLAT —The BLAST -Like Alignment Tool. Kent (2002). *Genome Research*. doi:10.1101/gr.229202

  > Analyzing vertebrate genomes requires rapid mRNA/DNA and cross-species protein alignments. A new tool, BLAT, is more accurate and 500 times faster than popular existing tools for mRNA/DNA alignments and 50 times faster for protein alignments at sensitivity settings typically used when comparing vertebrate sequences. BLAT's speed stems from an index of all nonoverlapping K-mers in the genome. This index fits inside the RAM of inexpensive computers, and need only be computed once for each genome assembly. BLAT has several major stages. It uses the index to find regions in the genome likely to be homologous to the query sequence. It performs an alignment between homologous regions. It stitches together these aligned regions (often exons) into larger alignments (typically genes). Finally, BLAT revisits small internal exons possibly missed at the first stage and adjusts large gap boundaries that have canonical splice sites where feasible. This paper describes how BLAT was optimized. Effects on speed and sensitivity are explored for various K-mer sizes, mismatch schemes, and number of required index matches. BLAT is compared with other alignment programs on various test sets and then used in several genome-wide applications. http://genome.ucsc.edu hosts a web-based BLAT server for the human genome.

- Buchfink et al. (2021) — DIAMOND, BLAST-level sensitivity at much higher speed.

  Sensitive protein alignments at tree-of-life scale using DIAMOND. Buchfink et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01101-x

  > We are at the beginning of a genomic revolution in which all known species are planned to be sequenced. Accessing such data for comparative analyses is crucial in this new age of data-driven biology. Here, we introduce an improved version of DIAMOND that greatly exceeds previous search performances and harnesses supercomputing to perform tree-of-life scale protein alignments in hours, while matching the sensitivity of the gold standard BLASTP.

- Steinegger and Söding (2018) — Linclust: clustering huge sequence sets in linear time.

  Clustering huge protein sequence sets in linear time. Steinegger & Söding (2018). *Nature Communications*. doi:10.1038/s41467-018-04964-5

  > Metagenomic datasets contain billions of protein sequences that could greatly enhance large-scale functional annotation and structure prediction. Utilizing this enormous resource would require reducing its redundancy by similarity clustering. However, clustering hundreds of millions of sequences is impractical using current algorithms because their runtimes scale as the input set size N times the number of clusters K, which is typically of similar order as N, resulting in runtimes that increase almost quadratically with N. We developed Linclust, the first clustering algorithm whose runtime scales as N, independent of K. It can also cluster datasets several times larger than the available main memory. We cluster 1.6 billion metagenomic sequence fragments in 10 h on a single server to 50% sequence identity, >1000 times faster than has been possible before. Linclust will help to unlock the great wealth contained in metagenomic and genomic sequence databases.

- Barrio-Hernandez et al. (2023) — clustering the AlphaFold database with Foldseek.

  Clustering predicted structures at the scale of the known protein universe. Barrio-Hernandez et al. (2023). *Nature*. doi:10.1038/s41586-023-06510-w

  > Proteins are key to all cellular processes and their structure is important in understanding their function and evolution. Sequence-based predictions of protein structures have increased in accuracy 1 , and over 214 million predicted structures are available in the AlphaFold database 2 . However, studying protein structures at this scale requires highly efficient methods. Here, we developed a structural-alignment-based clustering algorithm-Foldseek cluster-that can cluster hundreds of millions of structures. Using this method, we have clustered all of the structures in the AlphaFold database, identifying 2.30 million non-singleton structural clusters, of which 31% lack annotations representing probable previously undescribed structures. Clusters without annotation tend to have few representatives covering only 4% of all proteins in the AlphaFold database. Evolutionary analysis suggests that most clusters are ancient in origin but 4% seem to be species specific, representing lower-quality predictions or examples of de novo gene birth. We also show how structural comparisons can be used to predict domain families and their relationships, identifying examples of remote structural similarity. On the basis of these analyses, we identify several examples of human immune-related proteins with putative remote homology in prokaryotic species, illustrating the value of this resource for studying protein function and evolution across the tree of life.

- Ferragina and Manzini (2000) — the FM-index.

  Opportunistic data structures with applications. Ferragina & Manzini (2000). *Proceedings 41st Annual Symposium on Foundations of Computer Science*. doi:10.1109/sfcs.2000.892127

  > We address the issue of compressing and indexing data. We devise a data structure whose space occupancy is a function of the entropy of the underlying data set. We call the data structure opportunistic since its space occupancy is decreased when the input is compressible and this space reduction is achieved at no significant slowdown in the query performance. More precisely, its space occupancy is optimal in an information-content sense because text T[1,u] is stored using O(H/sub k/(T))+o(1) bits per input symbol in the worst case, where H/sub k/(T) is the kth order empirical entropy of T (the bound holds for any fixed k). Given an arbitrary string P[1,p], the opportunistic data structure allows to search for the occurrences of P in T in O(p+occlog/sup /spl epsiv//u) time (for any fixed /spl epsiv/>0). If data are uncompressible we achieve the best space bound currently known (Grossi and Vitter, 2000); on compressible data our solution improves the succinct suffix array of (Grossi and Vitter, 2000) and the classical suffix tree and suffix array data structures either in space or in query time or both. We also study our opportunistic data structure in a dynamic setting and devise a variant achieving effective search and update time bounds. Finally, we show how to plug our opportunistic data structure into the Glimpse tool (Manber and Wu, 1994). The result is an indexing tool which achieves sublinear space and sublinear query time complexity.

B.1.5 Multiple sequence alignment and motifs

- Feng and Doolittle (1987) — progressive alignment.

  Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Feng & Doolittle (1987). *Journal of Molecular Evolution*. doi:10.1007/bf02603120

  > A progressive alignment method is described that utilizes the Needleman and Wunsch pairwise alignment algorithm iteratively to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The sequences are assumed a priori to share a common ancestor, and the trees are constructed from difference matrices derived directly from the multiple alignment. The thrust of the method involves putting more trust in the comparison of recently diverged sequences than in those evolved in the distant past. In particular, this rule is followed: "once a gap, always a gap." The method has been applied to three sets of protein sequences: 7 superoxide dismutases, 11 globins, and 9 tyrosine kinase-like sequences. Multiple alignments and phylogenetic trees for these sets of sequences were determined and compared with trees derived by conventional pairwise treatments. In several instances, the progressive method led to trees that appeared to be more in line with biological expectations than were trees obtained by more commonly used methods.

- Thompson et al. (1994) — ClustalW.

  CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Thompson et al. (1994). *Nucleic Acids Research*. doi:10.1093/nar/22.22.4673

  > The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Firstly, individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Fourthly, positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions. These modifications are incorporated into a new program, CLUSTAL W which is freely available.

- Sievers et al. (2011) — Clustal Omega.

  Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Sievers et al. (2011). *Molecular Systems Biology*. doi:10.1038/msb.2011.75

  > Multiple sequence alignments are fundamental to many sequence analysis methods. Most alignments are computed using the progressive alignment heuristic. These methods are starting to become a bottleneck in some analysis pipelines when faced with data sets of the size of many thousands of sequences. Some methods allow computation of larger data sets while sacrificing quality, and others produce high-quality alignments, but scale badly with the number of sequences. In this paper, we describe a new program called Clustal Omega, which can align virtually any number of protein sequences quickly and that delivers accurate alignments. The accuracy of the package on smaller test cases is similar to that of the high-quality aligners. On larger data sets, Clustal Omega outperforms other packages in terms of execution time and quality. Clustal Omega also has powerful features for adding sequences to and exploiting information in existing alignments, making use of the vast amount of precomputed information in public databases like Pfam.

- Notredame et al. (2000) — T-Coffee (consistency).

  T-coffee: a novel method for fast and accurate multiple sequence alignment. Notredame et al. (2000). *Journal of Molecular Biology*. doi:10.1006/jmbi.2000.4042

  > We describe a new method (T-Coffee) for multiple sequence alignment that provides a dramatic improvement in accuracy with a modest sacrifice in speed as compared to the most commonly used alternatives. The method is broadly based on the popular progressive approach to multiple alignment but avoids the most serious pitfalls caused by the greedy nature of this algorithm. With T-Coffee we pre-process a data set of all pair-wise alignments between the sequences. This provides us with a library of alignment information that can be used to guide the progressive alignment. Intermediate alignments are then based not only on the sequences to be aligned next but also on how all of the sequences align with each other. This alignment information can be derived from heterogeneous sources such as a mixture of alignment programs and/or structure superposition. Here, we illustrate the power of the approach by using a combination of local and global pair-wise alignments to generate the library. The resulting alignments are significantly more reliable, as determined by comparison with a set of 141 test cases, than any of the popular alternatives that we tried. The improvement, especially clear with the more difficult test cases, is always visible, regardless of the phylogenetic spread of the sequences in the tests.

- Edgar (2004) — MUSCLE.

  MUSCLE: multiple sequence alignment with high accuracy and high throughput. Edgar (2004). *Nucleic Acids Research*. doi:10.1093/nar/gkh340

  > We describe MUSCLE, a new computer program for creating multiple alignments of protein sequences. Elements of the algorithm include fast distance estimation using kmer counting, progressive alignment using a new profile function we call the log-expectation score, and refinement using tree-dependent restricted partitioning. The speed and accuracy of MUSCLE are compared with T-Coffee, MAFFT and CLUSTALW on four test sets of reference alignments: BAliBASE, SABmark, SMART and a new benchmark, PREFAB. MUSCLE achieves the highest, or joint highest, rank in accuracy on each of these sets. Without

> refinement, MUSCLE achieves average accuracy statistically indistinguishable from T-Coffee and MAFFT, and is the fastest of the tested methods for large numbers of sequences, aligning 5000 sequences of average length 350 in 7 min on a current desktop computer. The MUSCLE program, source code and PREFAB test data are freely available at http://www.drive5. com/muscle.

- Katoh and Standley (2013) — MAFFT.

  MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Katoh & Standley (2013). *Molecular Biology and Evolution*. doi:10.1093/molbev/mst010

  > We report a major update of the MAFFT multiple sequence alignment program. This version has several new features, including options for adding unaligned sequences into an existing alignment, adjustment of direction in nucleotide alignment, constrained alignment and parallel processing, which were implemented after the previous major update. This report shows actual examples to explain how these features work, alone and in combination. Some examples incorrectly aligned by MAFFT are also shown to clarify its limitations. We discuss how to avoid misalignments, and our ongoing efforts to overcome such limitations.

- Lawrence et al. (1993) — Gibbs sampling for subtle motifs.

  Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment. Lawrence et al. (1993). *Science*. doi:10.1126/science.8211139

  > A wealth of protein and DNA sequence data is being generated by genome projects and other sequencing efforts. A crucial barrier to deciphering these sequences and understanding the relations among them is the difficulty of detecting subtle local residue patterns common to multiple sequences. Such patterns frequently reflect similar molecular structures and biological properties. A mathematical definition of this "local multiple alignment" problem suitable for full computer automation has been used to develop a new and sensitive algorithm, based on the statistical method of iterative sampling. This algorithm finds an optimized local alignment model for N sequences in N-linear time, requiring only seconds on current workstations, and allows the simultaneous detection and optimization of multiple patterns and pattern repeats. The method is illustrated as applied to helix-turn-helix proteins, lipocalins, and prenyltransferases.

- Bailey and Elkan (1994) — MEME (expectation maximisation).

  Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Bailey & Elkan (1994). *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

- Schneider and Stephens (1990) — sequence logos.

  Sequence logos: a new way to display consensus sequences. Schneider & Stephens (1990). *Nucleic Acids Research*. doi:10.1093/nar/18.20.6097

  > A graphical method is presented for displaying the patterns in a set of aligned sequences. The characters representing the sequence are stacked on top of each other for each position in the aligned sequences. The height of each letter is made proportional to its frequency, and the letters are sorted so the most common one is on top. The height of the entire stack is then adjusted to signify the information content of the sequences at that position. From these 'sequence logos', one can determine not only the consensus sequence but also the relative frequency of bases and the information content (measured in bits) at every position in a site or sequence. The logo displays both significant residues and subtle sequence patterns.

B.1.6 Profiles and hidden Markov models

- Gribskov et al. (1987) — profile analysis.

  Profile analysis: detection of distantly related proteins.. Gribskov et al. (1987). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.84.13.4355

> Profile analysis is a method for detecting distantly related proteins by sequence comparison. The basis for comparison is not only the customary Dayhoff mutational-distance matrix but also the results of structural studies and information implicit in the alignments of the sequences of families of similar proteins. This information is expressed in a position-specific scoring table (profile), which is created from a group of sequences previously aligned by structural or sequence similarity. The similarity of any other sequence (target) to the group of aligned sequences (probe) can be tested by comparing the target to the profile using dynamic programming algorithms. The profile method differs in two major respects from methods of sequence comparison in common use: (i) Any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target than is possible with pairwise alignment methods. (ii) The profile includes the penalties for insertion or deletion at each position, which allow one to include the probe secondary structure in the testing scheme. Tests with globin and immunoglobulin sequences show that profile analysis can distinguish all members of these families from all other sequences in a database containing 3800 protein sequences.

- Park et al. (1998) — profile methods find three times as many remote homologues as pairwise methods.

  Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. Park et al. (1998). *Journal of Molecular Biology*. doi:10.1006/jmbi.1998.2221

  > The sequences of related proteins can diverge beyond the point where their relationship can be recognised by pairwise sequence comparisons. In attempts to overcome this limitation, methods have been developed that use as a query, not a single sequence, but sets of related sequences or a representation of the characteristics shared by related sequences. Here we describe an assessment of three of these methods: the SAM-T98 implementation of a hidden Markov model procedure; PSI-BLAST; and the intermediate sequence search (ISS) procedure. We determined the extent to which these procedures can detect evolutionary relationships between the members of the sequence database PDBD40-J. This database, derived from the structural classification of proteins (SCOP), contains the sequences of proteins of known structure whose sequence identities with each other are 40% or less. The evolutionary relationships that exist between those that have low sequence identities were found by the examination of their structural details and, in many cases, their functional features. For nine false positive predictions out of a possible 432,680, i.e. at a false positive rate of about 1/50,000, SAM-T98 found 35% of the true homologous relationships in PDBD40-J, whilst PSI-BLAST found 30% and ISS found 25%. Overall, this is about twice the number of PDBD40-J relations that can be detected by the pairwise comparison procedures FASTA (17%) and GAP-BLAST (15%). For distantly related sequences in PDBD40-J, those pairs whose sequence identity is less than 30%, SAM-T98 and PSI-BLAST detect three times the number of relationships found by the pairwise methods.

- Krogh et al. (1994) — profile HMMs for protein families.

  Hidden Markov Models in Computational Biology. Krogh et al. (1994). *Journal of Molecular Biology*. doi:10.1006/jmbi.1994.1104

  > Hidden Markov Models (HMMs) are applied to the problems of statistical modeling, database searching and multiple sequence alignment of protein families and protein domains. These methods are demonstrated on the globin family, the protein kinase catalytic domain, and the EF-hand calcium binding motif. In each case the parameters of an HMM are estimated from a training set of unaligned sequences. After the HMM is built, it is used to obtain a multiple alignment of all the training sequences. It is also used to search the SWISS-PROT 22 database for other sequences that are members of the given protein family, or contain the given domain. The HMM produces multiple alignments of good quality that agree closely with the alignments produced by programs that incorporate three-dimensional structural information. When employed in discrimination tests (by examining how closely the sequences in a database fit the globin, kinase and EF-hand HMMs), the HMM is able to distinguish members of these families from non-members with a high degree of accuracy. Both the HMM and PROFILESEARCH (a technique used to search for relationships between a protein sequence and multiply aligned sequences) perform better in these tests than PROSITE (a dictionary of sites and patterns in proteins). The HMM appears to have a slight advantage over PROFILESEARCH in terms of lower rates of false negatives and false positives, even though the HMM is trained using only unaligned sequences, whereas PROFILESEARCH requires aligned training sequences. Our results suggest the presence of an EF-hand calcium binding motif in a highly conserved and evolutionary preserved putative intracellular region of 155 residues in the alpha-1 subunit of L-type calcium channels which play an important role in excitation-contraction coupling. This

region has been suggested to contain the functional domains that are typical or essential for all L-type calcium channels regardless of whether they couple to ryanodine receptors, conduct ions or both.

- Rabiner (1989) — a tutorial on hidden Markov models.

  A tutorial on hidden Markov models and selected applications in speech recognition. Rabiner (1989). *Proceedings of the IEEE*. doi:10.1109/5.18626

  > This tutorial provides an overview of the basic theory of hidden Markov models (HMMs) as originated by L.E. Baum and T. Petrie (1966) and gives practical details on methods of implementation of the theory along with a description of selected applications of the theory to distinct problems in speech recognition. Results from a number of original sources are combined to provide a single source of acquiring the background required to pursue further this area of research. The author first reviews the theory of discrete Markov chains and shows how the concept of hidden states, where the observation is a probabilistic function of the state, can be used effectively. The theory is illustrated with two simple examples, namely coin-tossing, and the classic balls-in-urns system. Three fundamental problems of HMMs are noted and several practical techniques for solving these problems are given. The various types of HMMs that have been studied, including ergodic as well as left-right models, are described.>

- Eddy (2008) — a probabilistic model of local alignment and significance estimation.

  A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. Eddy (2008). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000069

  > Sequence database searches require accurate estimation of the statistical significance of scores. Optimal local sequence alignment scores follow Gumbel distributions, but determining an important parameter of the distribution (lambda) requires time-consuming computational simulation. Moreover, optimal alignment scores are less powerful than probabilistic scores that integrate over alignment uncertainty ("Forward" scores), but the expected distribution of Forward scores remains unknown. Here, I conjecture that both expected score distributions have simple, predictable forms when full probabilistic modeling methods are used. For a probabilistic model of local sequence alignment, optimal alignment bit scores ("Viterbi" scores) are Gumbel-distributed with constant lambda = log 2, and the high scoring tail of Forward scores is exponential with the same constant lambda. Simulation studies support these conjectures over a wide range of profile/sequence comparisons, using 9,318 profile-hidden Markov models from the Pfam database. This enables efficient and accurate determination of expectation values (E-values) for both Viterbi and Forward scores for probabilistic local alignments.

- Söding (2004) — HMM–HMM comparison (HHsearch).

  Protein homology detection by HMM–HMM comparison. Söding (2004). *Bioinformatics*. doi:10.1093/bioinformatics/bti125

  > Motivation Protein homology detection and sequence alignment are at the basis of protein structure prediction, function prediction and evolution. Results We have generalized the alignment of protein sequences with a profile hidden Markov model (HMM) to the case of pairwise alignment of profile HMMs. We present a method for detecting distant homologous relationships between proteins based on this approach. The method (HHsearch) is benchmarked together with BLAST, PSI-BLAST, HMMER and the profile-profile comparison tools PROF_SIM and COMPASS, in an all-against-all comparison of a database of 3691 protein domains from SCOP 1.63 with pairwise sequence identities below 20%.Sensitivity: When the predicted secondary structure is included in the HMMs, HHsearch is able to detect between 2.7 and 4.2 times more homologs than PSI-BLAST or HMMER and between 1.44 and 1.9 times more than COMPASS or PROF_SIM for a rate of false positives of 10%. Approximately half of the improvement over the profile-profile comparison methods is attributable to the use of profile HMMs in place of simple profiles. Alignment quality: Higher sensitivity is mirrored by an increased alignment quality. HHsearch produced 1.2, 1.7 and 3.3 times more good alignments ('balanced' score >0.3) than the next best method (COMPASS), and 1.6, 2.9 and 9.4 times more than PSI-BLAST, at the family, superfamily and fold level, respectively.Speed: HHsearch scans a query of 200 residues against 3691 domains in 33 s on an AMD64 2GHz PC. This is 10 times faster than PROF_SIM and 17 times faster than COMPASS.

- ⋆ Remmert et al. (2011) — HHblits.

  HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Remmert et al. (2011). *Nature Methods*. doi:10.1038/nmeth.1818

> Sequence-based protein function and structure prediction depends crucially on sequence-search sensitivity and accuracy of the resulting sequence alignments. We present an open-source, general-purpose tool that represents both query and database sequences by profile hidden Markov models (HMMs): 'HMM-HMM-based lightning-fast iterative sequence search' (HHblits; http://toolkit.genzentrum.lmu.de/hhblits/). Compared to the sequence-search tool PSI-BLAST, HHblits is faster owing to its discretized-profile prefilter, has 50-100% higher sensitivity and generates more accurate alignments.

- Steinegger et al. (2019) — HH-suite3.

  HH-suite3 for fast remote homology detection and deep protein annotation. Steinegger et al. (2019). *BMC Bioinformatics*. doi:10.1186/s12859-019-3019-7

  > Background HH-suite is a widely used open source software suite for sensitive sequence similarity searches and protein fold recognition. It is based on pairwise alignment of profile Hidden Markov models (HMMs), which represent multiple sequence alignments of homologous proteins. Results We developed a single-instruction multiple-data (SIMD) vectorized implementation of the Viterbi algorithm for profile HMM alignment and introduced various other speed-ups. These accelerated the search methods HHsearch by a factor 4 and HHblits by a factor 2 over the previous version 2.0.16. HHblits3 is ~10× faster than PSI-BLAST and ~20× faster than HMMER3. Jobs to perform HHsearch and HHblits searches with many query profile HMMs can be parallelized over cores and over cluster servers using OpenMP and message passing interface (MPI). The free, open-source, GPLv3-licensed software is available at https://github.com/soedinglab/hh-suite . Conclusion The added functionalities and increased speed of HHsearch and HHblits should facilitate their use in large-scale protein structure and function prediction, e.g. in metagenomics and genomics projects.

B.1.7 RNA homology search

- Eddy and Durbin (1994) — covariance models for RNA sequence analysis.

  RNA sequence analysis using covariance models. Eddy & Durbin (1994). *Nucleic Acids Research*. doi:10.1093/nar/22.11.2079

  > We describe a general approach to several RNA sequence analysis problems using probabilistic models that flexibly describe the secondary structure and primary sequence consensus of an RNA sequence family. We call these models 'covariance models'. A covariance model of tRNA sequences is an extremely sensitive and discriminative tool for searching for additional tRNAs and tRNA-related sequences in sequence databases. A model can be built automatically from an existing sequence alignment. We also describe an algorithm for learning a model and hence a consensus secondary structure from initially unaligned example sequences and no prior structural information. Models trained on unaligned tRNA examples correctly predict tRNA secondary structure and produce high-quality multiple alignments. The approach may be applied to any family of small RNA sequences.

- Nawrocki and Eddy (2013) — Infernal 1.1: 100-fold faster RNA homology searches.

  Infernal 1.1: 100-fold faster RNA homology searches. Nawrocki & Eddy (2013). *Bioinformatics*. doi:10.1093/bioinformatics/btt509

  > Infernal builds probabilistic profiles of the sequence and secondary structure of an RNA family called covariance models (CMs) from structurally annotated multiple sequence alignments given as input. Infernal uses CMs to search for new family members in sequence databases and to create potentially large multiple sequence alignments. Version 1.1 of Infernal introduces a new filter pipeline for RNA homology search based on accelerated profile hidden Markov model (HMM) methods and HMM-banded CM alignment methods. This enables ~100-fold acceleration over the previous version and ~10 000-fold acceleration over exhaustive non-filtered CM searches. Availability Source code, documentation and the benchmark are downloadable from http://infernal.janelia.org. Infernal is freely licensed under the GNU GPLv3 and should be portable to any POSIX-compliant operating system, including Linux and Mac OS/X. Documentation includes a user's guide with a tutorial, a discussion of file formats and user options and additional details on methods implemented in the software. Contact nawrockie@janelia.hhmi.org

- Griffiths-Jones et al. (2003), Kalvari et al. (2020) — the Rfam database.

  Rfam: an RNA family database. Griffiths-Jones et al. (2003). *Nucleic Acids Research*. doi:10.1093/nar/gkg006

  > Rfam is a collection of multiple sequence alignments and covariance models representing non-coding RNA families. Rfam is available on the web in the UK at http://www.sanger.ac.uk/Software/Rfam/ and in the US at http://rfam.wustl.edu/. These websites allow the user to search a query sequence against a library of covariance models, and view multiple sequence alignments and family annotation. The database can also be downloaded in flatfile form and searched locally using the INFERNAL package (http://infernal.wustl.edu/). The first release of Rfam (1.0) contains 25 families, which annotate over 50 000 non-coding RNA genes in the taxonomic divisions of the EMBL nucleotide database.

  Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Kalvari et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa1047

  > Rfam is a database of RNA families where each of the 3444 families is represented by a multiple sequence alignment of known RNA sequences and a covariance model that can be used to search for additional members of the family. Recent developments have involved expert collaborations to improve the quality and coverage of Rfam data, focusing on microRNAs, viral and bacterial RNAs. We have completed the first phase of synchronising microRNA families in Rfam and miRBase, creating 356 new Rfam families and updating 40. We established a procedure for comprehensive annotation of viral RNA families starting with Flavivirus and Coronaviridae RNAs. We have also increased the coverage of bacterial and metagenome-based RNA families from the ZWD database. These developments have enabled a significant growth of the database, with the addition of 759 new families in Rfam 14. To facilitate further community contribution to Rfam, expert users are now able to build and submit new families using the newly developed Rfam Cloud family curation system. New Rfam website features include a new sequence similarity search powered by RNAcentral, as well as search and visualisation of families with pseudoknots. Rfam is freely available at https://rfam.org.

- Nussinov and Jacobson (1980), Zuker and Stiegler (1981) — RNA secondary structure by dynamic programming.

  Fast algorithm for predicting the secondary structure of single-stranded RNA.. Nussinov & Jacobson (1980). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.77.11.6309

  > A computer method is presented for finding the most stable secondary structures in long single-stranded RNAs. It is 1-2 orders of magnitude faster than existing codes. The time required for its application increases as N3 for a chain N nucleotides long. As many as 1000 nucleotides can be searched in a single run. The approach is systematic and builds an optimal structure in a straightforward inductive procedure based on an exact mathematical algorithm. Two simple half-matrices are constructed and the best folded form is read directly from the second matrix by a simple back-tracking procedure. The program utilizes published values for base-pairing energies to compute one structure with the lowest free energy.

  Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Zuker & Stiegler (1981). *Nucleic Acids Research*. doi:10.1093/nar/9.1.133

  > This paper presents a new computer method for folding an RNA molecule that finds a conformation of minimum free energy using published values of stacking and destabilizing energies. It is based on a dynamic programming algorithm from applied mathematics, and is much more efficient, faster, and can fold larger molecules than procedures which have appeared up to now in the biological literature. Its power is demonstrated in the folding of a 459 nucleotide immunoglobulin gamma 1 heavy chain messenger RNA fragment. We go beyond the basic method to show how to incorporate additional information into the algorithm. This includes data on chemical reactivity and enzyme susceptibility. We illustrate this with the folding of two large fragments from the 16S ribosomal RNA of Escherichia coli.

- Sankoff (1985) — simultaneous folding and alignment.

  Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. Sankoff (1985). *SIAM Journal on Applied Mathematics*. doi:10.1137/0145048

> The alignment of finite sequences, the inference of ribonucleic acid secondary structures (folding), and the reconstruction of ancestral sequences on a phylogenetic tree, are three problems which have dynamic programming solutions, which we formulate in a common mathematical framework. Combining the objective functions for alignment (parsimony, or minimal mutations) and folding (free energy), we present an algorithm which solves all three problems simultaneously for a set of N sequences of length n in time proportional to $n^{3N} $ and storage $n^{2N} $. Incorporating a "cutting corners" constraint against biologically unlikely alignments reduces these requirements so that they are proportional to $n^{3} $ and $n^{2} $, respectively, for fixed N.

- Rivas et al. (2016) — a statistical test for conserved RNA structure.

  A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Rivas et al. (2016). *Nature Methods*. doi:10.1038/nmeth.4066

  > Many functional RNAs have an evolutionarily conserved secondary structure. Conservation of RNA base pairing induces pairwise covariations in sequence alignments. We developed a computational method, R-scape (RNA Structural Covariation Above Phylogenetic Expectation), that quantitatively tests whether covariation analysis supports the presence of a conserved RNA secondary structure. R-scape analysis finds no statistically significant support for proposed secondary structures of the long noncoding RNAs HOTAIR, SRA, and Xist.

- Lowe and Eddy (1997) — tRNAscan-SE.

  tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Lowe & Eddy (1997). *Nucleic Acids Research*. doi:10.1093/nar/25.5.955

  > We describe a program, tRNAscan-SE, which identifies 99-100% of transfer RNA genes in DNA sequence while giving less than one false positive per 15 gigabases. Two previously described tRNA detection programs are used as fast, first-pass prefilters to identify candidate tRNAs, which are then analyzed by a highly selective tRNA covariance model. This work represents a practical application of RNA covariance models, which are general, probabilistic secondary structure profiles based on stochastic context-free grammars. tRNAscan-SE searches at approximately 30 000 bp/s. Additional extensions to tRNAscan-SE detect unusual tRNA homologues such as selenocysteine tRNAs, tRNA-derived repetitive elements and tRNA pseudogenes.

B.1.8 Alignment-free comparison

- Zielezinski et al. (2017) — review of alignment-free sequence comparison.

  Alignment-free sequence comparison: benefits, applications, and tools. Zielezinski et al. (2017). *Genome Biology*. doi:10.1186/s13059-017-1319-7

  > Alignment-free sequence analyses have been applied to problems ranging from whole-genome phylogeny to the classification of protein families, identification of horizontally transferred genes, and detection of recombined sequences. The strength of these methods makes them particularly useful for next-generation sequencing data processing and analysis. However, many researchers are unclear about how these methods work, how they compare to alignment-based methods, and what their potential is for use for their research. We address these questions and provide a guide to the currently available alignment-free sequence analysis tools.

- Ondov et al. (2016) — Mash: genome distances with MinHash.

  Mash: fast genome and metagenome distance estimation using MinHash. Ondov et al. (2016). *Genome Biology*. doi:10.1186/s13059-016-0997-x

  > Mash extends the MinHash dimensionality-reduction technique to include a pairwise mutation distance and P value significance test, enabling the efficient clustering and search of massive sequence collections. Mash reduces large sequences and sequence sets to small, representative sketches, from which global mutation distances can be rapidly estimated. We demonstrate several use cases, including the clustering of all 54,118 NCBI RefSeq genomes in 33 CPU h; real-time database search using assembled or unassembled Illumina, Pacific Biosciences, and Oxford Nanopore data; and the scalable clustering

of hundreds of metagenomic samples by composition. Mash is freely released under a BSD license (https://github.com/marbl/mash).

- Wood and Salzberg (2014) — Kraken: metagenomic classification with *k*-mers.

  Kraken: ultrafast metagenomic sequence classification using exact alignments. Wood & Salzberg (2014). *Genome Biology*. doi:10.1186/gb-2014-15-3-r46

  > Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences. Previous programs designed for this task have been relatively slow and computationally expensive, forcing researchers to use faster abundance estimation programs, which only classify small subsets of metagenomic data. Using exact alignment of k-mers, Kraken achieves classification accuracy comparable to the fastest BLAST program. In its fastest mode, Kraken classifies 100 base pair reads at a rate of over 4.1 million reads per minute, 909 times faster than Megablast and 11 times faster than the abundance estimation program MetaPhlAn. Kraken is available at http://ccb.jhu.edu/software/kraken/.

- C. Jain et al. (2018) — ANI and prokaryotic species boundaries.

  High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Jain et al. (2018). *Nature Communications*. doi:10.1038/s41467-018-07641-9

  > A fundamental question in microbiology is whether there is continuum of genetic diversity among genomes, or clear species boundaries prevail instead. Whole-genome similarity metrics such as Average Nucleotide Identity (ANI) help address this question by facilitating high resolution taxonomic analysis of thousands of genomes from diverse phylogenetic lineages. To scale to available genomes and beyond, we present FastANI, a new method to estimate ANI using alignment-free approximate sequence mapping. FastANI is accurate for both finished and draft genomes, and is up to three orders of magnitude faster compared to alignment-based approaches. We leverage FastANI to compute pairwise ANI values among all prokaryotic genomes available in the NCBI database. Our results reveal clear genetic discontinuity, with 99.8% of the total 8 billion genome pairs analyzed conforming to >95% intra-species and <83% inter-species ANI values. This discontinuity is manifested with or without the most frequently sequenced species, and is robust to historic additions in the genome databases.

B.1.9 Benchmarking

- Brenner et al. (1998) — assessing sequence comparison methods with structurally identified distant relationships.

  Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Brenner et al. (1998). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.95.11.6073

  > Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) J. Mol. Biol. 247, 536-540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). J. Mol. Biol. 215, 403-410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) Methods Enzymol. 266, 460-480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444-2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) J. Mol. Biol. 147, 195-197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20-30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

- Lindahl and Elofsson (2000) — identification of related proteins at family, superfamily and fold level.

  Identification of related proteins on family, superfamily and fold level. Lindahl & Elofsson (2000). *Journal of Molecular Biology*. doi:10.1006/jmbi.1999.3377

  > Proteins might have considerable structural similarities even when no evolutionary relationship of their sequences can be detected. This property is often referred to as the proteins sharing only a "fold". Of course, there are also sequences of common origin in each fold, called a "superfamily", and in them groups of sequences with clear similarities, designated "family". Developing algorithms to reliably identify proteins related at any level is one of the most important challenges in the fast growing field of bioinformatics today. However, it is not at all certain that a method proficient at finding sequence similarities performs well at the other levels, or vice versa.Here, we have compared the performance of various search methods on these different levels of similarity. As expected, we show that it becomes much harder to detect proteins as their sequences diverge. For family related sequences the best method gets 75% of the top hits correct. When the sequences differ but the proteins belong to the same superfamily this drops to 29%, and in the case of proteins with only fold similarity it is as low as 15%. We have made a more complete analysis of the performance of different algorithms than earlier studies, also including threading methods in the comparison. Using this method a more detailed picture emerges, showing multiple sequence information to improve detection on the two closer levels of relationship. We have also compared the different methods of including this information in prediction algorithms. For lower specificities, the best scheme to use is a linking method connecting proteins through an intermediate hit. For higher specificities, better performance is obtained by PSI-BLAST and some procedures using hidden Markov models. We also show that a threading method, THREADER, performs significantly better than any other method at fold recognition.

- Gribskov and Robinson (1996) — ROC analysis of sequence matching.

  Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Gribskov & Robinson (1996). *Computers & Chemistry*. doi:10.1016/s0097-8485(96)80004-0

  > In this paper, we borrow the idea of the receiver operating characteristic (ROC) from clinical medicine and demonstrate its application to sequence comparison. The ROC includes elements of both sensitivity and specificity, and is a quantitative measure of the usefulness of a diagnostic. The ROC is used in this work to investigate the effects of scoring table and gap penalties on database searches. Studies on three families of proteins, 4Fe-4S ferredoxins, lysR bacterial regulatory proteins, and bacterial RNA polymerase sigma-factors lead to the following conclusions: sequence families are quite idiosyncratic, but the best PAM distance for database searches using the Smith-Waterman method is somewhat larger than predicted by theoretical methods, about 200 PAM. The length independent gap penalty (gap initiation penalty) is quite important, but shows a broad peak at values of about 20-24. The length dependent gap penalty (gap extension penalty) is almost irrelevant suggesting that successful database searches rely only to a limited degree on gapped alignments. Taken together, these observations lead to the conclusion that the optimal conditions for alignments and database searches are not, and should not be expected to be, the same.

B.1.10 Domains

- Sonnhammer et al. (1997) — Pfam: domain families from seed alignments.

  Pfam: A comprehensive database of protein domain families based on seed alignments. Sonnhammer et al. (1997). *Proteins: Structure, Function, and Genetics*. doi:10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

  > Databases of multiple sequence alignments are a valuable aid to protein sequence classification and analysis. One of the main challenges when constructing such a database is to simultaneously satisfy the conflicting demands of completeness on the one hand and quality of alignment and domain definitions on the other. The latter properties are best dealt with by manual approaches, whereas completeness in practice is only amenable to automatic methods. Herein we present a database based on hidden Markov model profiles (HMMs), which combines high quality and completeness. Our database, Pfam, consists of parts A and B. Pfam-A is curated and contains well-characterized protein domain families with high quality alignments, which are maintained by using manually checked seed alignments and HMMs to find and align all members. Pfam-B contains sequence families that were generated automatically by applying

> the Domainer algorithm to cluster and align the remaining protein sequences after removal of Pfam-A domains. By using Pfam, a large number of previously unannotated proteins from the Caenorhabditis elegans genome project were classified. We have also identified many novel family memberships in known proteins, including new kazal, Fibronectin type III, and response regulator receiver domains. Pfam-A families have permanent accession numbers and form a library of HMMs available for searching and automatic annotation of new protein sequences.

- Mistry et al. (2020), Paysan-Lafosse et al. (2024) — Pfam today, including Pfam-N.

  Pfam: The protein families database in 2021. Mistry et al. (2020). *Nucleic Acids Research.* doi:10.1093/nar/gkaa913

  > The Pfam database is a widely used resource for classifying protein sequences into families and domains. Since Pfam was last described in this journal, over 350 new families have been added in Pfam 33.1 and numerous improvements have been made to existing entries. To facilitate research on COVID-19, we have revised the Pfam entries that cover the SARS-CoV-2 proteome, and built new entries for regions that were not covered by Pfam. We have reintroduced Pfam-B which provides an automatically generated supplement to Pfam and contains 136 730 novel clusters of sequences that are not yet matched by a Pfam family. The new Pfam-B is based on a clustering by the MMseqs2 software. We have compared all of the regions in the RepeatsDB to those in Pfam and have started to use the results to build and refine Pfam repeat families. Pfam is freely available for browsing and download at http://pfam.xfam.org/.

  The Pfam protein families database: embracing AI/ML. Paysan-Lafosse et al. (2024). *Nucleic Acids Research.* doi:10.1093/nar/gkae997

  > The Pfam protein families database is a comprehensive collection of protein domains and families used for genome annotation and protein structure and function analysis (https://www.ebi.ac.uk/interpro/). This update describes major developments in Pfam since 2020, including decommissioning the Pfam website and integration with InterPro, harmonization with the ECOD structural classification, and expanded curation of metagenomic, microprotein and repeat-containing families. We highlight how AlphaFold structure predictions are being leveraged to refine domain boundaries and identify new domains. New families discovered through large-scale sequence similarity analysis of AlphaFold models are described. We also detail the development of Pfam-N, which uses deep learning to expand family coverage, achieving an 8.8% increase in UniProtKB coverage compared to standard Pfam. We discuss plans for more frequent Pfam releases integrated with InterPro and the potential for artificial intelligence to further assist curation. Despite recent advances, many protein families remain to be classified, and Pfam continues working toward comprehensive coverage of the protein universe.

- Bileschi et al. (2022) — deep learning to annotate the protein universe (Pfam-N).

  Using deep learning to annotate the protein universe. Bileschi et al. (2022). *Nature Biotechnology.* doi:10.1038/s41587-021-01179-w

  > Understanding the relationship between amino acid sequence and protein function is a long-standing challenge with far-reaching scientific and translational implications. State-of-the-art alignment-based techniques cannot predict function for one-third of microbial protein sequences, hampering our ability to exploit data from diverse organisms. Here, we train deep learning models to accurately predict functional annotations for unaligned amino acid sequences across rigorous benchmark assessments built from the 17,929 families of the protein families database Pfam. The models infer known patterns of evolutionary substitutions and learn representations that accurately cluster sequences from unseen families. Combining deep models with existing methods significantly improves remote homology detection, suggesting that the deep models learn complementary information. This approach extends the coverage of Pfam by >9.5%, exceeding additions made over the last decade, and predicts function for 360 human reference proteome proteins with no previous Pfam annotation. These results suggest that deep learning models will be a core component of future protein annotation tools.

- Lau et al. (2024) — TED: the Encyclopedia of Domains in the AlphaFold database.

  Exploring structural diversity across the protein universe with The Encyclopedia of Domains. Lau et al. (2024). *Science.* doi:10.1126/science.adq4946

> The AlphaFold Protein Structure Database (AFDB) contains more than 214 million predicted protein structures composed of domains, which are independently folding units found in multiple structural and functional contexts. Identifying domains can enable many functional and evolutionary analyses but has remained challenging because of the sheer scale of the data. Using deep learning methods, we have detected and classified every domain in the AFDB, producing The Encyclopedia of Domains. We detected nearly 365 million domains, over 100 million more than can be found by sequence methods, covering more than 1 million taxa. Reassuringly, 77% of the nonredundant domains are similar to known superfamilies, greatly expanding representation of their domain space. We uncovered more than 10,000 new structural interactions between superfamilies and thousands of new folds across the fold space continuum.

B.1.11 Protein language models

- Rives et al. (2021) — ESM-1b: biological structure and function emerge from unsupervised learning.

  Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Rives et al. (2021). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.2016239118

  > In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In the life sciences, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Protein language modeling at the scale of evolution is a logical step toward predictive and generative artificial intelligence for biology. To this end, we use unsupervised learning to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity. The resulting model contains information about biological properties in its representations. The representations are learned from sequence data alone. The learned representation space has a multiscale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins. Information about secondary and tertiary structure is encoded in the representations and can be identified by linear projections. Representation learning produces features that generalize across a range of applications, enabling state-of-the-art supervised prediction of mutational effect and secondary structure and improving state-of-the-art features for long-range contact prediction.

- Elnaggar et al. (2022) — ProtTrans.

  ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. Elnaggar et al. (2022). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/tpami.2021.3095381

  > Computational biology and bioinformatics provide vast data gold-mines from protein sequences, ideal for Language Models (LMs) taken from Natural Language Processing (NLP). These LMs reach for new prediction frontiers at low inference costs. Here, we trained two auto-regressive models (Transformer-XL, XLNet) and four auto-encoder models (BERT, Albert, Electra, T5) on data from UniRef and BFD containing up to 393 billion amino acids. The protein LMs (pLMs) were trained on the Summit supercomputer using 5616 GPUs and TPU Pod up-to 1024 cores. Dimensionality reduction revealed that the raw pLM-embeddings from unlabeled data captured some biophysical features of protein sequences. We validated the advantage of using the embeddings as exclusive input for several subsequent tasks: (1) a per-residue (per-token) prediction of protein secondary structure (3-state accuracy Q3=81%-87%); (2) per-protein (pooling) predictions of protein sub-cellular location (ten-state accuracy: Q10=81%) and membrane versus water-soluble (2-state accuracy Q2=91%). For secondary structure, the most informative embeddings (ProtT5) for the first time outperformed the state-of-the-art without multiple sequence alignments (MSAs) or evolutionary information thereby bypassing expensive database searches. Taken together, the results implied that pLMs learned some of the grammar of the language of life. All our models are available through https://github.com/agemagician/ProtTrans.

- Lin et al. (2023) — ESM-2 and ESMFold.

  Evolutionary-scale prediction of atomic-level protein structure with a language model. Lin et al. (2023). *Science*. doi:10.1126/science.ade2574

> Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

- Hamamsy et al. (2023) — TM-Vec: remote homology from sequence embeddings.

  Protein remote homology detection and structural alignment using deep learning. Hamamsy et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-023-01917-2

  > Exploiting sequence-structure-function relationships in biotechnology requires improved methods for aligning proteins that have low sequence similarity to previously annotated proteins. We develop two deep learning methods to address this gap, TM-Vec and DeepBLAST. TM-Vec allows searching for structure-structure similarities in large sequence databases. It is trained to accurately predict TM-scores as a metric of structural similarity directly from sequence pairs without the need for intermediate computation or solution of structures. Once structurally similar proteins have been identified, DeepBLAST can structurally align proteins using only sequence information by identifying structurally homologous regions between proteins. It outperforms traditional sequence alignment methods and performs similarly to structure-based alignment methods. We show the merits of TM-Vec and DeepBLAST on a variety of datasets, including better identification of remotely homologous proteins compared with state-of-the-art sequence alignment and structure prediction methods.

- Hong et al. (2024) — dense retrieval for fast, sensitive homologue detection.

  Fast, sensitive detection of protein homologs using deep dense retrieval. Hong et al. (2024). *Nature Biotechnology*. doi:10.1038/s41587-024-02353-6

  > The identification of protein homologs in large databases using conventional methods, such as protein sequence comparison, often misses remote homologs. Here, we offer an ultrafast, highly sensitive method, dense homolog retriever (DHR), for detecting homologs on the basis of a protein language model and dense retrieval techniques. Its dual-encoder architecture generates different embeddings for the same protein sequence and easily locates homologs by comparing these representations. Its alignment-free nature improves speed and the protein language model incorporates rich evolutionary and structural information within DHR embeddings. DHR achieves a >10% increase in sensitivity compared to previous methods and a >56% increase in sensitivity at the superfamily level for samples that are challenging to identify using alignment-based approaches. It is up to 22 times faster than traditional methods such as PSI-BLAST and DIAMOND and up to 28,700 times faster than HMMER. The new remote homologs exclusively found by DHR are useful for revealing connections between well-characterized proteins and improving our knowledge of protein evolution, structure and function.

- Kaminski et al. (2023) — pLM-BLAST.

  pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. Kaminski et al. (2023). *Bioinformatics*. doi:10.1093/bioinformatics/btad579

  > Motivation The detection of homology through sequence comparison is a typical first step in the study of protein function and evolution. In this work, we explore the applicability of protein language models to this task. Results We introduce pLM-BLAST, a tool inspired by BLAST, that detects distant homology by comparing single-sequence representations (embeddings) derived from a protein language model, ProtT5. Our benchmarks reveal that pLM-BLAST maintains a level of accuracy on par with HHsearch for both highly similar sequences (with >50% identity) and markedly divergent sequences (with Availability and implementation pLM-BLAST is accessible via the MPI Bioinformatics Toolkit as a web server for searching precomputed databases (https://toolkit.tuebingen.mpg.de/tools/plmblast). It is also available as a standalone tool for building custom databases and performing batch searches (https://github.com/labstructbioinf/pLM-BLAST).

- Liu et al. (2024) — PLMSearch.

  PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. Liu et al. (2024). *Nature Communications*. doi:10.1038/s41467-024-46808-5

  > Homologous protein search is one of the most commonly used methods for protein annotation and analysis. Compared to structure search, detecting distant evolutionary relationships from sequences alone remains challenging. Here we propose PLMSearch (Protein Language Model), a homologous protein search method with only sequences as input. PLMSearch uses deep representations from a pre-trained protein language model and trains the similarity prediction model with a large number of real structure similarity. This enables PLMSearch to capture the remote homology information concealed behind the sequences. Extensive experimental results show that PLMSearch can search millions of query-target protein pairs in seconds like MMseqs2 while increasing the sensitivity by more than threefold, and is comparable to state-of-the-art structure search methods. In particular, unlike traditional sequence search methods, PLMSearch can recall most remote homology pairs with dissimilar sequences but similar structures. PLMSearch is freely available at https://dmiip.sjtu.edu.cn/PLMSearch .

B.2 Databases

Further papers for Chapter 2.

In the HTML version, click *Abstract* to show the abstract.

B.2.1 General

- Rigden and Fernández (2025) — the 2026 NAR database issue.

  The 2026 Nucleic Acids Research database issue and the online molecular biology database collection. Rigden & Fernández (2025). *Nucleic Acids Research*. doi:10.1093/nar/gkaf1427

  > The 2026 Nucleic Acids Research database issue has 182 papers from across biology and neighbouring fields. Eighty-four of these papers describe new databases, while 86 are updates on databases that have previously appeared here. Twelve more papers cover databases most recently published elsewhere. New nucleic acid databases include NapRNAdb for noncapped RNA and GlycoRNAdb. Protein structure is covered by updates from wwPDB members and the AlphaFold Database; SMART, PROSITE, and eggNOG cover domains and families. The Open Enzyme Database and QSproteome are new community-orientated initiatives. JoGo covers hierarchically named and contextualised human haplotypes in the issue's first Breakthrough paper; So3D provides genuinely 3D spatial transcriptomics in the other. Foundational databases Genenames.org and Gene Ontology also provide updates. The Database Issue is freely available on the Nucleic Acids Research website (https://academic.oup.com/nar). At the NAR online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/c/), over the past year, 899 entries were reviewed, 96 new resources added, and 319 discontinued URLs removed, bringing the total number of databases to 2173.

- Sayers et al. (2024) — database resources of NCBI.

  Database resources of the National Center for Biotechnology Information in 2025. Sayers et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae979

  > The National Center for Biotechnology Information (NCBI) provides online information resources for biology, including the GenBank® nucleic acid sequence repository and the PubMed® repository of citations and abstracts published in life science journals. NCBI provides search and retrieval operations for most of these data from 31 distinct repositories and knowledgebases. The E-utilities serve as the programming interface for most of these. Resources receiving significant updates in the past year include PubMed, PubMed Central, Bookshelf, the NIH Comparative Genomics Resource, BLAST, Sequence Read Archive, Taxonomy, iCn3D, Conserved Domain Database, Pathogen Detection, antimicrobial resistance resources and PubChem. These resources can be accessed through the NCBI home page at https://www.ncbi.nlm.nih.gov.

B.2.2 Sequences and genomes

- Arita et al. (2020) — the International Nucleotide Sequence Database Collaboration.

  The international nucleotide sequence database collaboration. Arita et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa967

  > The International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/) has been the core infrastructure for collecting and providing nucleotide sequence data and metadata for >30 years. Three partner organizations, the DNA Data Bank of Japan (DDBJ) at the National Institute of Genetics in Mishima, Japan; the European Nucleotide Archive (ENA) at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK; and GenBank at National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health in Bethesda, Maryland, USA have been collaboratively maintaining the INSDC for the benefit of not only science but all types of community worldwide.

- O'Leary et al. (2015) — RefSeq.

  Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. O'Leary et al. (2015). *Nucleic Acids Research*. doi:10.1093/nar/gkv1189

  > The RefSeq project at the National Center for Biotechnology Information (NCBI) maintains and curates a publicly available database of annotated genomic, transcript, and protein sequence records (http://www.ncbi.nlm.nih.gov/refseq/). The RefSeq project leverages the data submitted to the International Nucleotide Sequence Database Collaboration (INSDC) against a combination of computation, manual curation, and collaboration to produce a standard set of stable, non-redundant reference sequences. The RefSeq project augments these reference sequences with current knowledge including publications, functional features and informative nomenclature. The database currently represents sequences from more than 55,000 organisms (>4800 viruses, >40,000 prokaryotes and >10,000 eukaryotes; RefSeq release 71), ranging from a single record to complete genomes. This paper summarizes the current status of the viral, prokaryotic, and eukaryotic branches of the RefSeq project, reports on improvements to data access and details efforts to further expand the taxonomic representation of the collection. We also highlight diverse functional curation initiatives that support multiple uses of RefSeq data including taxonomic validation, genome annotation, comparative genomics, and clinical testing. We summarize our approach to utilizing available RNA-Seq and other data types in our manual curation process for vertebrate, plant, and other species, and describe a new direction for prokaryotic genomes and protein name management.

- Suzek et al. (2014) — UniRef clusters.

  UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Suzek et al. (2014). *Bioinformatics*. doi:10.1093/bioinformatics/btu739

  > Motivation UniRef databases provide full-scale clustering of UniProtKB sequences and are utilized for a broad range of applications, particularly similarity-based functional annotation. Non-redundancy and intra-cluster homogeneity in UniRef were recently improved by adding a sequence length overlap threshold. Our hypothesis is that these improvements would enhance the speed and sensitivity of similarity searches and improve the consistency of annotation within clusters. Results Intra-cluster molecular function consistency was examined by analysis of Gene Ontology terms. Results show that UniRef clusters bring together proteins of identical molecular function in more than 97% of the clusters, implying that clusters are useful for annotation and can also be used to detect annotation inconsistencies. To examine coverage in similarity results, BLASTP searches against UniRef50 followed by expansion of the hit lists with cluster members demonstrated advantages compared with searches against UniProtKB sequences; the searches are concise (~7 times shorter hit list before expansion), faster (~6 times) and more sensitive in detection of remote similarities (>96% recall at e-value <0.0001). Our results support the use of UniRef clusters as a comprehensive and scalable alternative to native sequence databases for similarity searches and reinforces its reliability for use in functional annotation.

- Richardson et al. (2022) — MGnify.

  MGnify: the microbiome sequence data analysis resource in 2023. Richardson et al. (2022). *Nucleic Acids Research*. doi:10.1093/nar/gkac1080

> The MGnify platform (https://www.ebi.ac.uk/metagenomics) facilitates the assembly, analysis and archiving of microbiome-derived nucleic acid sequences. The platform provides access to taxonomic assignments and functional annotations for nearly half a million analyses covering metabarcoding, metatranscriptomic, and metagenomic datasets, which are derived from a wide range of different environments. Over the past 3 years, MGnify has not only grown in terms of the number of datasets contained but also increased the breadth of analyses provided, such as the analysis of long-read sequences. The MGnify protein database now exceeds 2.4 billion non-redundant sequences predicted from metagenomic assemblies. This collection is now organised into a relational database making it possible to understand the genomic context of the protein through navigation back to the source assembly and sample metadata, marking a major improvement. To extend beyond the functional annotations already provided in MGnify, we have applied deep learning-based annotation methods. The technology underlying MGnify's Application Programming Interface (API) and website has been upgraded, and we have enabled the ability to perform downstream analysis of the MGnify data through the introduction of a coupled Jupyter Lab environment.

- Dyer et al. (2024) — Ensembl 2025.

  Ensembl 2025. Dyer et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1071

  > Ensembl (www.ensembl.org) is an open platform integrating publicly available genomics data across the tree of life with a focus on eukaryotic species related to human health, agriculture and biodiversity. This year has seen a continued expansion in the number of species represented, with >4800 eukaryotic and >31 300 prokaryotic genomes available. The new Ensembl site, currently in beta, has continued to develop, currently holding >2700 eukaryotic genome assemblies. The new site provides genome, gene, transcript, homology and variation views, and will replace the current Rapid Release site; this represents a key step towards provision of a single integrated Ensembl site. Additional activities have included developing improved regulatory annotation for human, mouse and agricultural species, and expanding the Ensembl Variant Effect Predictor tool. To learn more about Ensembl, help and documentation are available along with an extensive training program that can be accessed via our training pages.

- Perez et al. (2024) — the UCSC Genome Browser.

  The UCSC Genome Browser database: 2025 update. Perez et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae974

  > The UCSC Genome Browser (https://genome.ucsc.edu) is a widely utilized web-based tool for visualization and analysis of genomic data, encompassing over 4000 assemblies from diverse organisms. Since its release in 2001, it has become an essential resource for genomics and bioinformatics research. Annotation data available on Genome Browser includes both internally created and maintained tracks as well as custom tracks and track hubs provided by the research community. This last year's updates include over 25 new annotation tracks such as the gnomAD 4.1 track on the human GRCh38/hg38 assembly, the addition of three new public hubs, and significant expansions to the Genome Archive[GenArk) system for interacting with the enormous variety of assemblies. We have also made improvements to our interface, including updates to the browser graphic page, such as a new popup dialog feature that now displays item details without requiring navigation away from the main Genome Browser page. GenePred tracks have been upgraded with right-click options for zooming and precise navigation, along with enhanced mouseOver functions. Additional improvements include a new grouping feature for track hubs and hub description info links. A new tutorial focusing on Clinical Genetics has also been added to the UCSC Genome Browser.

B.2.3 Protein families and structures

- Sonnhammer et al. (1997), Mistry et al. (2020) — Pfam.

  Pfam: A comprehensive database of protein domain families based on seed alignments. Sonnhammer et al. (1997). *Proteins: Structure, Function, and Genetics*. doi:10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

  > Databases of multiple sequence alignments are a valuable aid to protein sequence classification and analysis. One of the main challenges when constructing such a database is to simultaneously satisfy the conflicting demands of completeness on the one hand and quality of alignment and domain definitions on the other. The latter properties are best dealt with by manual approaches, whereas completeness in practice is only amenable to automatic methods. Herein we present a database based on hidden Markov

model profiles (HMMs), which combines high quality and completeness. Our database, Pfam, consists of parts A and B. Pfam-A is curated and contains well-characterized protein domain families with high quality alignments, which are maintained by using manually checked seed alignments and HMMs to find and align all members. Pfam-B contains sequence families that were generated automatically by applying the Domainer algorithm to cluster and align the remaining protein sequences after removal of Pfam-A domains. By using Pfam, a large number of previously unannotated proteins from the Caenorhabditis elegans genome project were classified. We have also identified many novel family memberships in known proteins, including new kazal, Fibronectin type III, and response regulator receiver domains. Pfam-A families have permanent accession numbers and form a library of HMMs available for searching and automatic annotation of new protein sequences.

Pfam: The protein families database in 2021. Mistry et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa913

> The Pfam database is a widely used resource for classifying protein sequences into families and domains. Since Pfam was last described in this journal, over 350 new families have been added in Pfam 33.1 and numerous improvements have been made to existing entries. To facilitate research on COVID-19, we have revised the Pfam entries that cover the SARS-CoV-2 proteome, and built new entries for regions that were not covered by Pfam. We have reintroduced Pfam-B which provides an automatically generated supplement to Pfam and contains 136 730 novel clusters of sequences that are not yet matched by a Pfam family. The new Pfam-B is based on a clustering by the MMseqs2 software. We have compared all of the regions in the RepeatsDB to those in Pfam and have started to use the results to build and refine Pfam repeat families. Pfam is freely available for browsing and download at http://pfam.xfam.org/.

- Berman et al. (2000) — the Protein Data Bank.

  The Protein Data Bank. Berman et al. (2000). *Nucleic Acids Research*. doi:10.1093/nar/28.1.235

  > The Protein Data Bank (PDB; http://www.rcsb.org/pdb/) is the single worldwide archive of structural data of biological macromolecules. This paper describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and near-term plans for the future development of the resource.

- Turner et al. (2023) — the Electron Microscopy Data Bank.

  EMDB—the Electron Microscopy Data Bank. Turner et al. (2023). *Nucleic Acids Research*. doi:10.1093/nar/gkad1019

  > The Electron Microscopy Data Bank (EMDB) is the global public archive of three-dimensional electron microscopy (3DEM) maps of biological specimens derived from transmission electron microscopy experiments. As of 2021, EMDB is managed by the Worldwide Protein Data Bank consortium (wwPDB; wwpdb.org) as a wwPDB Core Archive, and the EMDB team is a core member of the consortium. Today, EMDB houses over 30 000 entries with maps containing macromolecules, complexes, viruses, organelles and cells. Herein, we provide an overview of the rapidly growing EMDB archive, including its current holdings, recent updates, and future plans.

- Murzin et al. (1995), Sillitoe et al. (2020), Cheng et al. (2014), Lau et al. (2024) — SCOP, CATH, ECOD and TED.

  SCOP: A structural classification of proteins database for the investigation of sequences and structures. Murzin et al. (1995). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(05)80134-2

  CATH: increased structural coverage of functional space. Sillitoe et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa1079

  > CATH (https://www.cathdb.info) identifies domains in protein structures from wwPDB and classifies these into evolutionary superfamilies, thereby providing structural and functional annotations. There are two levels: CATH-B, a daily snapshot of the latest domain structures and superfamily assignments, and CATH+, with additional derived data, such as predicted sequence domains, and functionally coherent sequence subsets (Functional Families or FunFams). The latest CATH+ release, version 4.3, significantly increases coverage of structural and sequence data, with an addition of 65,351 fully-classified domains structures (+15%), providing 500 238 structural domains, and 151 million predicted sequence domains (+59%) assigned

to 5481 superfamilies. The FunFam generation pipeline has been re-engineered to cope with the increased influx of data. Three times more sequences are captured in FunFams, with a concomitant increase in functional purity, information content and structural coverage. FunFam expansion increases the structural annotations provided for experimental GO terms (+59%). We also present CATH-FunVar web-pages displaying variations in protein sequences and their proximity to known or predicted functional sites. We present two case studies (1) putative cancer drivers and (2) SARS-CoV-2 proteins. Finally, we have improved links to and from CATH including SCOP, InterPro, Aquaria and 2DProt.

ECOD: An Evolutionary Classification of Protein Domains. Cheng et al. (2014). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1003926

Understanding the evolution of a protein, including both close and distant relationships, often reveals insight into its structure and function. Fast and easy access to such up-to-date information facilitates research. We have developed a hierarchical evolutionary classification of all proteins with experimentally determined spatial structures, and presented it as an interactive and updatable online database. ECOD (Evolutionary Classification of protein Domains) is distinct from other structural classifications in that it groups domains primarily by evolutionary relationships (homology), rather than topology (or "fold"). This distinction highlights cases of homology between domains of differing topology to aid in understanding of protein structure evolution. ECOD uniquely emphasizes distantly related homologs that are difficult to detect, and thus catalogs the largest number of evolutionary links among structural domain classifications. Placing distant homologs together underscores the ancestral similarities of these proteins and draws attention to the most important regions of sequence and structure, as well as conserved functional sites. ECOD also recognizes closer sequence-based relationships between protein domains. Currently, approximately 100,000 protein structures are classified in ECOD into 9,000 sequence families clustered into close to 2,000 evolutionary groups. The classification is assisted by an automated pipeline that quickly and consistently classifies weekly releases of PDB structures and allows for continual updates. This synchronization with PDB uniquely distinguishes ECOD among all protein classifications. Finally, we present several case studies of homologous proteins not recorded in other classifications, illustrating the potential of how ECOD can be used to further biological and evolutionary studies.

Exploring structural diversity across the protein universe with The Encyclopedia of Domains. Lau et al. (2024). *Science*. doi:10.1126/science.adq4946

The AlphaFold Protein Structure Database (AFDB) contains more than 214 million predicted protein structures composed of domains, which are independently folding units found in multiple structural and functional contexts. Identifying domains can enable many functional and evolutionary analyses but has remained challenging because of the sheer scale of the data. Using deep learning methods, we have detected and classified every domain in the AFDB, producing The Encyclopedia of Domains. We detected nearly 365 million domains, over 100 million more than can be found by sequence methods, covering more than 1 million taxa. Reassuringly, 77% of the nonredundant domains are similar to known superfamilies, greatly expanding representation of their domain space. We uncovered more than 10,000 new structural interactions between superfamilies and thousands of new folds across the fold space continuum.

- Varadi et al. (2021), Varadi et al. (2023) — the AlphaFold database.

AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Varadi et al. (2021). *Nucleic Acids Research*. doi:10.1093/nar/gkab1061

The AlphaFold Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk) is an openly accessible, extensive database of high-accuracy protein-structure predictions. Powered by AlphaFold v2.0 of DeepMind, it has enabled an unprecedented expansion of the structural coverage of the known protein-sequence space. AlphaFold DB provides programmatic access to and interactive visualization of predicted atomic coordinates, per-residue and pairwise model-confidence estimates and predicted aligned errors. The initial release of AlphaFold DB contains over 360,000 predicted structures across 21 model-organism proteomes, which will soon be expanded to cover most of the (over 100 million) representative sequences from the UniRef90 data set.

AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Varadi et al. (2023). *Nucleic Acids Research*. doi:10.1093/nar/gkad1011

> The AlphaFold Database Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk) has significantly impacted structural biology by amassing over 214 million predicted protein structures, expanding from the initial 300k structures released in 2021. Enabled by the groundbreaking AlphaFold2 artificial intelligence (AI) system, the predictions archived in AlphaFold DB have been integrated into primary data resources such as PDB, UniProt, Ensembl, InterPro and MobiDB. Our manuscript details subsequent enhancements in data archiving, covering successive releases encompassing model organisms, global health proteomes, Swiss-Prot integration, and a host of curated protein datasets. We detail the data access mechanisms of AlphaFold DB, from direct file access via FTP to advanced queries using Google Cloud Public Datasets and the programmatic access endpoints of the database. We also discuss the improvements and services added since its initial release, including enhancements to the Predicted Aligned Error viewer, customisation options for the 3D viewer, and improvements in the search engine of AlphaFold DB.

- Lin et al. (2023) — ESMFold and the ESM Metagenomic Atlas.

  Evolutionary-scale prediction of atomic-level protein structure with a language model. Lin et al. (2023). *Science*. doi:10.1126/science.ade2574

  > Recent advances in machine learning have leveraged evolutionary information in multiple sequence alignments to predict protein structure. We demonstrate direct inference of full atomic-level protein structure from primary sequence using a large language model. As language models of protein sequences are scaled up to 15 billion parameters, an atomic-resolution picture of protein structure emerges in the learned representations. This results in an order-of-magnitude acceleration of high-resolution structure prediction, which enables large-scale structural characterization of metagenomic proteins. We apply this capability to construct the ESM Metagenomic Atlas by predicting structures for >617 million metagenomic protein sequences, including >225 million that are predicted with high confidence, which gives a view into the vast breadth and diversity of natural proteins.

- Kim et al. (2024) — BFVD, predicted viral protein structures.

  BFVD—a large repository of predicted viral protein structures. Kim et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1119

  > The AlphaFold Protein Structure Database (AFDB) is the largest repository of accurately predicted structures with taxonomic labels. Despite providing predictions for over 214 million UniProt entries, the AFDB does not cover viral sequences, severely limiting their study. To address this, we created the Big Fantastic Virus Database (BFVD), a repository of 351 242 protein structures predicted by applying ColabFold to the viral sequence representatives of the UniRef30 clusters. By utilizing homology searches across two petabases of assembled sequencing data, we improved 36% of these structure predictions beyond ColabFold's initial results. BFVD holds a unique repertoire of protein structures as over 62% of its entries show no or low structural similarity to existing repositories. We demonstrate how a substantial fraction of bacteriophage proteins, which remained unannotated based on their sequences, can be matched with similar structures from BFVD. In that, BFVD is on par with the AFDB, while holding nearly three orders of magnitude fewer structures. BFVD is an important virus-specific expansion to protein structure repositories, offering new opportunities to advance viral research. BFVD can be freely downloaded at bfvd.steineggerlab.workers.dev and queried using Foldseek and UniProt labels at bfvd.foldseek.com.

B.2.4 Function and pathways

- Ashburner et al. (2000) — the original Gene Ontology paper.

  Gene Ontology: tool for the unification of biology. Ashburner et al. (2000). *Nature Genetics*. doi:10.1038/75556

- Kanehisa et al. (2024) — KEGG.

  KEGG: biological systems database as a model of the real world. Kanehisa et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae909

> KEGG (https://www.kegg.jp/) is a database resource for representation and analysis of biological systems. Pathway maps are the primary dataset in KEGG representing systemic functions of the cell and the organism in terms of molecular interaction and reaction networks. The KEGG Orthology (KO) system is a mechanism for linking genes and proteins to pathway maps and other molecular networks. Each KO is a generic gene identifier and each pathway map is created as a network of KO nodes. This architecture enables KEGG pathway mapping to uncover systemic features from KO assigned genomes and metagenomes. Additional roles of KOs include characterization of conserved genes and conserved units of genes in organism groups, which can be done by taxonomy mapping. A new tool has been developed for identifying conserved gene orders in chromosomes, in which gene orders are treated as sequences of KOs. Furthermore, a new dataset called VOG (virus ortholog group) is computationally generated from virus proteins and expanded to proteins of cellular organisms, allowing gene orders to be compared as VOG sequences as well. Together with these datasets and analysis tools, new types of pathway maps are being developed to present a global view of biological processes involving multiple organism groups.

- Milacic et al. (2023) — Reactome.

  The Reactome Pathway Knowledgebase 2024. Milacic et al. (2023). *Nucleic Acids Research.* doi:10.1093/nar/gkad1025

  > The Reactome Knowledgebase (https://reactome.org), an Elixir and GCBR core biological data resource, provides manually curated molecular details of a broad range of normal and disease-related biological processes. Processes are annotated as an ordered network of molecular transformations in a single consistent data model. Reactome thus functions both as a digital archive of manually curated human biological processes and as a tool for discovering functional relationships in data such as gene expression profiles or somatic mutation catalogs from tumor cells. Here we review progress towards annotation of the entire human proteome, targeted annotation of disease-causing genetic variants of proteins and of small-molecule drugs in a pathway context, and towards supporting explicit annotation of cell- and tissue-specific pathways. Finally, we briefly discuss issues involved in making Reactome more fully interoperable with other related resources such as the Gene Ontology and maintaining the resulting community resource network.

- De Paolis Kaluza et al. (2026) — the fifth CAFA challenge.

  Advances in Protein Function Prediction from the Fifth CAFA Challenge. De Paolis Kaluza et al. (2026). *bioRxiv (preprint).* doi:10.64898/2026.04.27.716980

  > The Critical Assessment of Functional Annotation (CAFA) is a long-standing community effort to independently assess computational methods for protein function prediction, to highlight well-performing methodologies, to identify bottlenecks in the field, and to provide a forum for the dissemination of results and exchange of ideas. In its fifth round (CAFA5) of triennial challenges, a partnership with Kaggle Inc. facilitated participation from a large community of data scientists and computational biologists through a competitive prospective challenge on the crowdsourcing platform. In this work, we present an in-depth analysis of the submitted predictions and report improvements in accuracy over all methods from the previous CAFA challenges. We further introduce a new evaluation setting for proteins with pre-existing (incomplete) annotations and identify the need for methods that better leverage existing annotations to predict those that will be discovered later. Finally, we characterize the prospective evaluation framework by examining performance on a strict set of unpublished annotations and across intermediate database releases. Our results indicate that recent developments in the field, such as the availability of protein language models and accurately predicted 3D structures, as well as the growth of experimental annotations through biocuration, have all contributed to performance improvements. 1

B.2.5 Interactions

- Toro et al. (2021) — IntAct.

  The IntAct database: efficient access to fine-grained molecular interaction data. del Toro et al. (2021). *Nucleic Acids Research.* doi:10.1093/nar/gkab1006

> The IntAct molecular interaction database (https://www.ebi.ac.uk/intact) is a curated resource of molecular interactions, derived from the scientific literature and from direct data depositions. As of August 2021, IntAct provides more than one million binary interactions, curated by twelve global partners of the International Molecular Exchange consortium, for which the IntAct database provides a shared curation and dissemination platform. The IMEx curation policy has always emphasised a fine-grained data and curation model, aiming to capture the relevant experimental detail essential for the interpretation of the provided molecular interaction data. Here, we present recent curation focus and progress, as well as a completely redeveloped website which presents IntAct data in a much more user-friendly and detailed way.

- Oughtred et al. (2020) — BioGRID.

  The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Oughtred et al. (2020). *Protein Science*. doi:10.1002/pro.3978

  > The BioGRID (Biological General Repository for Interaction Datasets, thebiogrid.org) is an open-access database resource that houses manually curated protein and genetic interactions from multiple species including yeast, worm, fly, mouse, and human. The ~1.93 million curated interactions in BioGRID can be used to build complex networks to facilitate biomedical discoveries, particularly as related to human health and disease. All BioGRID content is curated from primary experimental evidence in the biomedical literature, and includes both focused low-throughput studies and large high-throughput datasets. BioGRID also captures protein post-translational modifications and protein or gene interactions with bioactive small molecules including many known drugs. A built-in network visualization tool combines all annotations and allows users to generate network graphs of protein, genetic and chemical interactions. In addition to general curation across species, BioGRID undertakes themed curation projects in specific aspects of cellular regulation, for example the ubiquitin-proteasome system, as well as specific disease areas, such as for the SARS-CoV-2 virus that causes COVID-19 severe acute respiratory syndrome. A recent extension of BioGRID, named the Open Repository of CRISPR Screens (ORCS, orcs.thebiogrid.org), captures single mutant phenotypes and genetic interactions from published high throughput genome-wide CRISPR/Cas9-based genetic screens. BioGRID-ORCS contains datasets for over 1,042 CRISPR screens carried out to date in human, mouse and fly cell lines. The biomedical research community can freely access all BioGRID data through the web interface, standardized file downloads, or via model organism databases and partner meta-databases.

- Meldal et al. (2021) — Complex Portal.

  Complex Portal 2022: new curation frontiers. Meldal et al. (2021). *Nucleic Acids Research*. doi:10.1093/nar/gkab991

  > The Complex Portal (www.ebi.ac.uk/complexportal) is a manually curated, encyclopaedic database of macromolecular complexes with known function from a range of model organisms. It summarizes complex composition, topology and function along with links to a large range of domain-specific resources (i.e. wwPDB, EMDB and Reactome). Since the last update in 2019, we have produced a first draft complexome for Escherichia coli, maintained and updated that of Saccharomyces cerevisiae, added over 40 coronavirus complexes and increased the human complexome to over 1100 complexes that include approximately 200 complexes that act as targets for viral proteins or are part of the immune system. The display of protein features in ComplexViewer has been improved and the participant table is now colour-coordinated with the nodes in ComplexViewer. Community collaboration has expanded, for example by contributing to an analysis of putative transcription cofactors and providing data accessible to semantic web tools through Wikidata which is now populated with manually curated Complex Portal content through a new bot. Our data license is now CC0 to encourage data reuse. Users are encouraged to get in touch, provide us with feedback and send curation requests through the 'Support' link.

- Buzzao, Persson, et al. (2024) — FunCoup 6.

  FunCoup 6: advancing functional association networks across species with directed links and improved user experience. Buzzao et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1021

  > FunCoup 6 (https://funcoup.org) represents a significant advancement in global functional association networks, aiming to provide researchers with a comprehensive view of the functional coupling interactome. This update introduces novel methodologies and integrated tools for improved network inference and analysis. Major new developments in FunCoup 6 include vastly expanding the coverage of gene regulatory

links, a new framework for bin-free Bayesian training and a new website. FunCoup 6 integrates a new tool for disease and drug target module identification using the TOPAS algorithm. To expand the utility of the resource for biomedical research, it incorporates pathway enrichment analysis using the ANUBIX and EASE algorithms. The unique comparative interactomics analysis in FunCoup provides insights of network conservation, now allowing users to align orthologs only or query each species network independently. Bin-free training was applied to 23 primary species, and in addition, networks were generated for all remaining 618 species in InParanoiDB 9. Accompanying these advancements, FunCoup 6 features a new redesigned website, together with updated API functionalities, and represents a pivotal step forward in functional genomics research, offering unique capabilities for exploring the complex landscape of protein interactions.

- Fischer et al. (2025) — hu.MAP 3.0.

  hu.MAP3.0: atlas of human protein complexes by integration of >25,000 proteomic experiments. Fischer et al. (2025). *Molecular Systems Biology*. doi:10.1038/s44320-025-00121-5

  Macromolecular protein complexes carry out most cellular functions. Unfortunately, we lack the subunit composition for many human protein complexes. To address this gap we integrated >25,000 mass spectrometry experiments using a machine learning approach to identify >15,000 human protein complexes. We show our map of protein complexes is highly accurate and more comprehensive than previous maps, placing nearly 70% of human proteins into their physical contexts. We globally characterize our complexes using mass spectrometry based protein covariation data (ProteomeHD.2) and identify covarying complexes suggesting common functional associations. hu.MAP3.0 generates testable functional hypotheses for 472 uncharacterized proteins which we support using AlphaFold modeling. Additionally, we use AlphaFold modeling to identify 5871 mutually exclusive proteins in hu.MAP3.0 complexes suggesting complexes serve different functional roles depending on their subunit composition. We identify expression as the primary way cells and organisms relieve the conflict of mutually exclusive subunits. Finally, we import our complexes to EMBL-EBI's Complex Portal (https://www.ebi.ac.uk/complexportal/home) and provide complexes through our hu.MAP3.0 web interface (https://humap3.proteincomplexes.org/). We expect our resource to be highly impactful to the broader research community.

B.2.6 Expression, proteomics and medical data

- Clough et al. (2023) — GEO.

  NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. Clough et al. (2023). *Nucleic Acids Research*. doi:10.1093/nar/gkad965

  The Gene Expression Omnibus (GEO) is an international public repository that archives gene expression and epigenomics data sets generated by next-generation sequencing and microarray technologies. Data are typically submitted to GEO by researchers in compliance with widespread journal and funder mandates to make generated data publicly accessible. The resource handles raw data files, processed data files and descriptive metadata for over 200 000 studies and 6.5 million samples, all of which are indexed, searchable and downloadable. Additionally, GEO offers web-based tools that facilitate analysis and visualization of differential gene expression. This article presents the current status and recent advancements in GEO, including the generation of consistently computed gene expression count matrices for thousands of RNA-seq studies, and new interactive graphical plots in GEO2R that help users identify differentially expressed genes and assess data set quality. The GEO repository is built and maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), and is publicly accessible at https://www.ncbi.nlm.nih.gov/geo/.

- Perez-Riverol et al. (2024) — PRIDE.

  The PRIDE database at 20 years: 2025 update. Perez-Riverol et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1011

  The PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) is the world's leading mass spectrometry (MS)-based proteomics data repository and one of the founding members of the ProteomeXchange consortium. This manuscript summarizes the developments in PRIDE resources and related tools for the last three years. The number of submitted datasets to PRIDE Archive (the archival component of PRIDE) has reached on average around 534 datasets per month. This has been possible thanks to continuous improvements in infrastructure such as a new file transfer protocol for very

large datasets (Globus), a new data resubmission pipeline and an automatic dataset validation process. Additionally, we will highlight novel activities such as the availability of the PRIDE chatbot (based on the use of open-source Large Language Models), and our work to improve support for MS crosslinking datasets. Furthermore, we will describe how we have increased our efforts to reuse, reanalyze and disseminate high-quality proteomics data into added-value resources such as UniProt, Ensembl and Expression Atlas.

- Uhlén et al. (2015) — the Human Protein Atlas.

  Tissue-based map of the human proteome. Uhlén et al. (2015). *Science*. doi:10.1126/science.1260419

  > Resolving the molecular details of proteome variation in the different tissues and organs of the human body will greatly increase our knowledge of human biology and disease. Here, we present a map of the human tissue proteome based on an integrated omics approach that involves quantitative transcriptomics at the tissue and organ level, combined with tissue microarray-based immunohistochemistry, to achieve spatial localization of proteins down to the single-cell level. Our tissue-based analysis detected more than 90% of the putative protein-coding genes. We used this approach to explore the human secretome, the membrane proteome, the druggable proteome, the cancer proteome, and the metabolic functions in 32 different tissues and organs. All the data are integrated in an interactive Web-based database that allows exploration of individual proteins, as well as navigation of global expression patterns, in all major tissues and organs in the human body.

- Chen et al. (2023) — gnomAD.

  A genomic mutational constraint map using variation in 76,156 human genomes. Chen et al. (2023). *Nature*. doi:10.1038/s41586-023-06045-0

  > The depletion of disruptive variation caused by purifying natural selection (constraint) has been widely used to investigate protein-coding genes underlying human disorders 1-4 , but attempts to assess constraint for non-protein-coding regions have proved more difficult. Here we aggregate, process and release a dataset of 76,156 human genomes from the Genome Aggregation Database (gnomAD)-the largest public open-access human genome allele frequency reference dataset-and use it to build a genomic constraint map for the whole genome (genomic non-coding constraint of haploinsufficient variation (Gnocchi)). We present a refined mutational model that incorporates local sequence context and regional genomic features to detect depletions of variation. As expected, the average constraint for protein-coding sequences is stronger than that for non-coding regions. Within the non-coding genome, constrained regions are enriched for known regulatory elements and variants that are implicated in complex human diseases and traits, facilitating the triangulation of biological annotation, disease association and natural selection to non-coding DNA analysis. More constrained regulatory elements tend to regulate more constrained protein-coding genes, which in turn suggests that non-coding constraint can aid the identification of constrained genes that are as yet unrecognized by current gene constraint metrics. We demonstrate that this genome-wide constraint map improves the identification and interpretation of functional human genetic variation.

- Landrum et al. (2017) — ClinVar.

  ClinVar: improving access to variant interpretations and supporting evidence. Landrum et al. (2017). *Nucleic Acids Research*. doi:10.1093/nar/gkx1153

  > ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) is a freely available, public archive of human genetic variants and interpretations of their significance to disease, maintained at the National Institutes of Health. Interpretations of the clinical significance of variants are submitted by clinical testing laboratories, research laboratories, expert panels and other groups. ClinVar aggregates data by variant-disease pairs, and by variant (or set of variants). Data aggregated by variant are accessible on the website, in an improved set of variant call format files and as a new comprehensive XML report. ClinVar recently started accepting submissions that are focused primarily on providing phenotypic information for individuals who have had genetic testing. Submissions may come from clinical providers providing their own interpretation of the variant ('provider interpretation') or from groups such as patient registries that primarily provide phenotypic information from patients ('phenotyping only'). ClinVar continues to make improvements to its search and retrieval functions. Several new fields are now indexed for more precise searching, and filters allow the user to narrow down a large set of search results.

- Amberger et al. (2018) — OMIM.

  OMIM.org: leveraging knowledge across phenotype–gene relationships. Amberger et al. (2018). *Nucleic Acids Research*. doi:10.1093/nar/gky1151

  > For over 50 years Mendelian Inheritance in Man has chronicled the collective knowledge of the field of medical genetics. It initially cataloged the known X-linked, autosomal recessive and autosomal dominant inherited disorders, but grew to be the primary repository of curated information on both genes and genetic phenotypes and the relationships between them. Each phenotype and gene is given a separate entry assigned a stable, unique identifier. The entries contain structured summaries of new and important information based on expert review of the biomedical literature. OMIM.org provides interactive access to the knowledge repository, including genomic coordinate searches of the gene map, views of genetic heterogeneity of phenotypes in Phenotypic Series, and side-by-side comparisons of clinical synopses. OMIM.org also supports computational queries via a robust API. All entries have extensive targeted links to other genomic resources and additional references. Updates to OMIM can be found on the update list or followed through the MIMmatch service. Updated user guides and tutorials are available on the website. As of September 2018, OMIM had over 24,600 entries, and the OMIM Morbid Map Scorecard had 6,259 molecularized phenotypes connected to 3,961 genes.

- Sollis et al. (2022) — the GWAS Catalog.

  The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Sollis et al. (2022). *Nucleic Acids Research*. doi:10.1093/nar/gkac1010

  > The NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas) is a FAIR knowledgebase providing detailed, structured, standardised and interoperable genome-wide association study (GWAS) data to >200 000 users per year from academic research, healthcare and industry. The Catalog contains variant-trait associations and supporting metadata for >45 000 published GWAS across >5000 human traits, and >40 000 full P-value summary statistics datasets. Content is curated from publications or acquired via author submission of prepublication summary statistics through a new submission portal and validation tool. GWAS data volume has vastly increased in recent years. We have updated our software to meet this scaling challenge and to enable rapid release of submitted summary statistics. The scope of the repository has expanded to include additional data types of high interest to the community, including sequencing-based GWAS, gene-based analyses and copy number variation analyses. Community outreach has increased the number of shared datasets from under-represented traits, e.g. cancer, and we continue to contribute to awareness of the lack of population diversity in GWAS. Interoperability of the Catalog has been enhanced through links to other resources including the Polygenic Score Catalog and the International Mouse Phenotyping Consortium, refinements to GWAS trait annotation, and the development of a standard format for GWAS data.

- Weinstein et al. (2013) — TCGA Pan-Cancer.

  The Cancer Genome Atlas Pan-Cancer analysis project. Weinstein et al. (2013). *Nature Genetics*. doi:10.1038/ng.2764

  > The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. The Pan-Cancer initiative compares the first 12 tumor types profiled by TCGA. Analysis of the molecular aberrations and their functional roles across tumor types will teach us how to extend therapies effective in one cancer type to others with a similar genomic profile.

- Sondka et al. (2023) — COSMIC.

  COSMIC: a curated database of somatic variants and clinical data for cancer. Sondka et al. (2023). *Nucleic Acids Research*. doi:10.1093/nar/gkad986

  > The Catalogue Of Somatic Mutations In Cancer (COSMIC), https://cancer.sanger.ac.uk/cosmic, is an expert-curated knowledgebase providing data on somatic variants in cancer, supported by a comprehensive suite of tools for interpreting genomic data, discerning the impact of somatic alterations on disease, and facilitating translational research. The catalogue is accessed and used by thousands of cancer researchers and clinicians daily, allowing them to quickly access information from an immense pool of data curated from over 29 thousand scientific publications and large studies. Within the last 4 years, COSMIC has

substantially expanded its utility by adding new resources: the Mutational Signatures catalogue, the Cancer Mutation Census, and Actionability. To improve data accessibility and interoperability, somatic variants have received stable genomic identifiers that are associated with their genomic coordinates in GRCh37 and GRCh38, and new export files with reduced data redundancy have been made available for download.

B.3 Evolution

Further papers for Chapter 3.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.3.1 The origin of life

- Miller (1953), Parker et al. (2011) — the Miller–Urey experiment and its reanalysis.

  A Production of Amino Acids Under Possible Primitive Earth Conditions. Miller (1953). *Science*. doi:10.1126/science.117.3046.528

  Primordial synthesis of amines and amino acids in a 1958 Miller H 2 S-rich spark discharge experiment. Parker et al. (2011). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1019191108

  > Archived samples from a previously unreported 1958 Stanley Miller electric discharge experiment containing hydrogen sulfide (H(2)S) were recently discovered and analyzed using high-performance liquid chromatography and time-of-flight mass spectrometry. We report here the detection and quantification of primary amine-containing compounds in the original sample residues, which were produced via spark discharge using a gaseous mixture of H(2)S, CH(4), NH(3), and CO(2). A total of 23 amino acids and 4 amines, including 7 organosulfur compounds, were detected in these samples. The major amino acids with chiral centers are racemic within the accuracy of the measurements, indicating that they are not contaminants introduced during sample storage. This experiment marks the first synthesis of sulfur amino acids from spark discharge experiments designed to imitate primordial environments. The relative yield of some amino acids, in particular the isomers of aminobutyric acid, are the highest ever found in a spark discharge experiment. The simulated primordial conditions used by Miller may serve as a model for early volcanic plume chemistry and provide insight to the possible roles such plumes may have played in abiotic organic synthesis. Additionally, the overall abundances of the synthesized amino acids in the presence of H(2)S are very similar to the abundances found in some carbonaceous meteorites, suggesting that H(2)S may have played an important role in prebiotic reactions in early solar system environments.

- Oró (1960) — prebiotic synthesis of adenine.

  Synthesis of adenine from ammonium cyanide. Oró (1960). *Biochemical and Biophysical Research Communications*. doi:10.1016/0006-291x(60)90138-8

- Kruger et al. (1982), Guerrier-Takada et al. (1983) — the discovery of ribozymes.

  Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Kruger et al. (1982). *Cell*. doi:10.1016/0092-8674(82)90414-7

  > In the macronuclear rRNA genes of Tetrahymena thermophila, a 413 bp intervening sequence (IVS) interrupts the 26S rRNA-coding region. A restriction fragment of the rDNA containing the IVS and portions of the adjacent rRNA sequences (exons) was inserted downstream from the lac UV5 promoter in a recombinant plasmid. Transcription of this template by purified Escherichia coli RNA polymerase in vitro produced a shortened version of the pre-rRNA, which was then deproteinized. When incubated with monovalent and divalent cations and a guanosine factor, this RNA underwent splicing. The reactions that were characterized included the precise excision of the IVS, attachment of guanosine to the 5' end of the IVS, covalent cyclization of the IVS and ligation of the exons. We conclude that splicing activity is intrinsic to the structure of the RNA, and that enzymes, small nuclear RNAs and folding of the pre-rRNA

> into an RNP are unnecessary for these reactions. We propose that the IVS portion of the RNA has several enzyme-like properties that enable it to break and reform phosphodiester bonds. The finding of autocatalytic rearrangements of RNA molecules has implications for the mechanism and the evolution of other reactions that involve RNA.

The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Guerrier-Takada et al. (1983). *Cell*. doi:10.1016/0092-8674(83)90117-4

> The RNA moieties of ribonuclease P purified from both E. coli (M1 RNA) and B. subtilis (P-RNA) can cleave tRNA precursor molecules in buffers containing either 60 mM Mg2+ or 10 mM Mg2+ plus 1 mM spermidine. The RNA acts as a true catalyst under these conditions whereas the protein moieties of the enzymes alone show no catalytic activity. However, in buffers containing 5-10 mM Mg2+ (in the absence of spermidine) both kinds of subunits are required for enzymatic activity, as shown previously. In the presence of low concentrations of Mg2+, in vitro, the RNA and protein subunits from one species can complement subunits from the other species in reconstitution experiments. When the precursor to E. coli 4.5S RNA is used as a substrate, only the enzyme complexes formed with M1 RNA from E. coli and the protein moieties from either bacterial species are active.

- Gilbert (1986) — the RNA world named.

Origin of life: The RNA world. Gilbert (1986). *Nature*. doi:10.1038/319618a0

- Joyce (2002), Orgel (2004), Higgs and Lehman (2014) — reviews of the RNA world.

The antiquity of RNA-based evolution. Joyce (2002). *Nature*. doi:10.1038/418214a

> All life that is known to exist on Earth today and all life for which there is evidence in the geological record seems to be of the same form--one based on DNA genomes and protein enzymes. Yet there are strong reasons to conclude that DNA- and protein-based life was preceded by a simpler life form based primarily on RNA. This earlier era is referred to as the 'RNA world', during which the genetic information resided in the sequence of RNA molecules and the phenotype derived from the catalytic properties of RNA.

Prebiotic Chemistry and the Origin of the RNA World. Orgel (2004). *Critical Reviews in Biochemistry and Molecular Biology*. doi:10.1080/10409230490460765

> The demonstration that ribosomal peptide synthesis is a ribozyme-catalyzed reaction makes it almost certain that there was once an RNA World. The central problem for origin-of-life studies, therefore, is to understand how a protein-free RNA World became established on the primitive Earth. We first review the literature on the prebiotic synthesis of the nucleotides, the nonenzymatic synthesis and copying of polynucleotides, and the selection of ribozyme catalysts of a kind that might have facilitated polynucleotide replication. This leads to a brief outline of the Molecular Biologists' Dream, an optimistic scenario for the origin of the RNA World. In the second part of the review we point out the many unresolved problems presented by the Molecular Biologists' Dream. This in turn leads to a discussion of genetic systems simpler than RNA that might have ”invented” RNA. Finally, we review studies of prebiotic membrane formation.

The RNA World: molecular cooperation at the origins of life. Higgs & Lehman (2014). *Nature Reviews Genetics*. doi:10.1038/nrg3841

> The RNA World concept posits that there was a period of time in primitive Earth's history - about 4 billion years ago - when the primary living substance was RNA or something chemically similar. In the past 50 years, this idea has gone from speculation to a prevailing idea. In this Review, we summarize the key logic behind the RNA World and describe some of the most important recent advances that have been made to support and expand this logic. We also discuss the ways in which molecular cooperation involving RNAs would facilitate the emergence and early evolution of life. The immediate future of RNA World research should be a very dynamic one.

- Ban et al. (2000) — the ribosome is a ribozyme.

The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. Ban et al. (2000). *Science*. doi:10.1126/science.289.5481.905

> The large ribosomal subunit catalyzes peptide bond formation and binds initiation, termination, and elongation factors. We have determined the crystal structure of the large ribosomal subunit from Haloarcula marismortui at 2.4 angstrom resolution, and it includes 2833 of the subunit's 3045 nucleotides and 27 of its 31 proteins. The domains of its RNAs all have irregular shapes and fit together in the ribosome like the pieces of a three-dimensional jigsaw puzzle to form a large, monolithic structure. Proteins are abundant everywhere on its surface except in the active site where peptide bond formation occurs and where it contacts the small subunit. Most of the proteins stabilize the structure by interacting with several RNA domains, often using idiosyncratically folded extensions that reach into the subunit's interior.

- Powner et al. (2009), Sutherland (2015) — prebiotic synthesis of nucleotides.

  Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Powner et al. (2009). *Nature*. doi:10.1038/nature08013

  > At some stage in the origin of life, an informational polymer must have arisen by purely chemical means. According to one version of the 'RNA world' hypothesis this polymer was RNA, but attempts to provide experimental support for this have failed. In particular, although there has been some success demonstrating that 'activated' ribonucleotides can polymerize to form RNA, it is far from obvious how such ribonucleotides could have formed from their constituent parts (ribose and nucleobases). Ribose is difficult to form selectively, and the addition of nucleobases to ribose is inefficient in the case of purines and does not occur at all in the case of the canonical pyrimidines. Here we show that activated pyrimidine ribonucleotides can be formed in a short sequence that bypasses free ribose and the nucleobases, and instead proceeds through arabinose amino-oxazoline and anhydronucleoside intermediates. The starting materials for the synthesis-cyanamide, cyanoacetylene, glycolaldehyde, glyceraldehyde and inorganic phosphate-are plausible prebiotic feedstock molecules, and the conditions of the synthesis are consistent with potential early-Earth geochemical models. Although inorganic phosphate is only incorporated into the nucleotides at a late stage of the sequence, its presence from the start is essential as it controls three reactions in the earlier stages by acting as a general acid/base catalyst, a nucleophilic catalyst, a pH buffer and a chemical buffer. For prebiotic reaction sequences, our results highlight the importance of working with mixed chemical systems in which reactants for a particular reaction step can also control other steps.

  The Origin of Life—Out of the Blue. Sutherland (2015). *Angewandte Chemie International Edition*. doi:10.1002/anie.201506585

  > Either to sustain autotrophy, or as a prelude to heterotrophy, organic synthesis from an environmentally available C1 feedstock molecule is crucial to the origin of life. Recent findings augment key literature results and suggest that hydrogen cyanide--"Blausäure"--was that feedstock.

- Lincoln and Joyce (2009) — self-sustained replication of an RNA enzyme.

  Self-Sustained Replication of an RNA Enzyme. Lincoln & Joyce (2009). *Science*. doi:10.1126/science.1167856

  > An RNA enzyme that catalyzes the RNA-templated joining of RNA was converted to a format whereby two enzymes catalyze each other's synthesis from a total of four oligonucleotide substrates. These cross-replicating RNA enzymes undergo self-sustained exponential amplification in the absence of proteins or other biological materials. Amplification occurs with a doubling time of about 1 hour and can be continued indefinitely. Populations of various cross-replicating enzymes were constructed and allowed to compete for a common pool of substrates, during which recombinant replicators arose and grew to dominate the population. These replicating RNA enzymes can serve as an experimental model of a genetic system. Many such model systems could be constructed, allowing different selective outcomes to be related to the underlying properties of the genetic system.

- Wächtershäuser (1988), Martin and Russell (2003) — metabolism first, and hydrothermal vents.

  Before enzymes and templates: theory of surface metabolism. Wächtershäuser (1988). *Microbiological Reviews*. doi:10.1128/mr.52.4.452-484.1988

  On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Martin & Russell (2003). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. doi:10.1098/rstb.2002.1183

> All life is organized as cells. Physical compartmentation from the environment and self-organization of self-contained redox reactions are the most conserved attributes of living things, hence inorganic matter with such attributes would be life's most likely forebear. We propose that life evolved in structured iron monosulphide precipitates in a seepage site hydrothermal mound at a redox, pH and temperature gradient between sulphide-rich hydrothermal fluid and iron(II)-containing waters of the Hadean ocean floor. The naturally arising, three-dimensional compartmentation observed within fossilized seepage-site metal sulphide precipitates indicates that these inorganic compartments were the precursors of cell walls and membranes found in free-living prokaryotes. The known capability of FeS and NiS to catalyse the synthesis of the acetyl-methylsulphide from carbon monoxide and methylsulphide, constituents of hydrothermal fluid, indicates that pre-biotic syntheses occurred at the inner surfaces of these metal-sulphide-walled compartments, which furthermore restrained reacted products from diffusion into the ocean, providing sufficient concentrations of reactants to forge the transition from geochemistry to biochemistry. The chemistry of what is known as the RNA-world could have taken place within these naturally forming, catalyticwalled compartments to give rise to replicating systems. Sufficient concentrations of precursors to support replication would have been synthesized in situ geochemically and biogeochemically, with FeS (and NiS) centres playing the central catalytic role. The universal ancestor we infer was not a free-living cell, but rather was confined to the naturally chemiosmotic, FeS compartments within which the synthesis of its constituents occurred. The first free-living cells are suggested to have been eubacterial and archaebacterial chemoautotrophs that emerged more than 3.8 Gyr ago from their inorganic confines. We propose that the emergence of these prokaryotic lineages from inorganic confines occurred independently, facilitated by the independent origins of membrane-lipid biosynthesis: isoprenoid ether membranes in the archaebacterial and fatty acid ester membranes in the eubacterial lineage. The eukaryotes, all of which are ancestrally heterotrophs and possess eubacterial lipids, are suggested to have arisen ca. 2 Gyr ago through symbiosis involving an autotrophic archaebacterial host and a heterotrophic eubacterial symbiont, the common ancestor of mitochondria and hydrogenosomes. The attributes shared by all prokaryotes are viewed as inheritances from their confined universal ancestor. The attributes that distinguish eubacteria and archaebacteria, yet are uniform within the groups, are viewed as relics of their phase of differentiation after divergence from the non-free-living universal ancestor and before the origin of the free-living chemoautotrophic lifestyle. The attributes shared by eukaryotes with eubacteria and archaebacteria, respectively, are viewed as inheritances via symbiosis. The attributes unique to eukaryotes are viewed as inventions specific to their lineage. The origin of the eukaryotic endomembrane system and nuclear membrane are suggested to be the fortuitous result of the expression of genes for eubacterial membrane lipid synthesis by an archaebacterial genetic apparatus in a compartment that was not fully prepared to accommodate such compounds, resulting in vesicles of eubacterial lipids that accumulated in the cytosol around their site of synthesis. Under these premises, the most ancient divide in the living world is that between eubacteria and archaebacteria, yet the steepest evolutionary grade is that between prokaryotes and eukaryotes.

- Bell et al. (2015), Dodd et al. (2017) — the earliest evidence for life.

Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. Bell et al. (2015). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1517557112

> Evidence of life on Earth is manifestly preserved in the rock record. However, the microfossil record only extends to ~ 3.5 billion years (Ga), the chemofossil record arguably to ~ 3.8 Ga, and the rock record to 4.0 Ga. Detrital zircons from Jack Hills, Western Australia range in age up to nearly 4.4 Ga. From a population of over 10,000 Jack Hills zircons, we identified one >3.8-Ga zircon that contains primary graphite inclusions. Here, we report carbon isotopic measurements on these inclusions in a concordant, 4.10 ± 0.01-Ga zircon. We interpret these inclusions as primary due to their enclosure in a crack-free host as shown by transmission X-ray microscopy and their crystal habit. Their δ(13)CPDB of -24 ± 5‰ is consistent with a biogenic origin and may be evidence that a terrestrial biosphere had emerged by 4.1 Ga, or ~ 300 My earlier than has been previously proposed.

Evidence for early life in Earth's oldest hydrothermal vent precipitates. Dodd et al. (2017). *Nature*. doi:10.1038/nature21377

> Although it is not known when or where life on Earth began, some of the earliest habitable environments may have been submarine-hydrothermal vents. Here we describe putative fossilized microorganisms that are at least 3,770 million and possibly 4,280 million years old in ferruginous sedimentary rocks, interpreted as seafloor-hydrothermal vent-related precipitates, from the Nuvvuagittuq belt in Quebec, Canada. These structures occur as micrometre-scale haematite tubes and filaments with morphologies and mineral assemblages similar to those of filamentous microorganisms from modern hydrothermal vent precipitates and analogous microfossils in younger rocks. The Nuvvuagittuq rocks contain isotopically

> light carbon in carbonate and carbonaceous material, which occurs as graphitic inclusions in diagenetic carbonate rosettes, apatite blades intergrown among carbonate rosettes and magnetite-haematite granules, and is associated with carbonate in direct contact with the putative microfossils. Collectively, these observations are consistent with an oxidized biomass and provide evidence for biological activity in submarine-hydrothermal environments more than 3,770 million years ago.

- Des Marais et al. (2008) — the NASA astrobiology roadmap.

  The NASA Astrobiology Roadmap. Des Marais et al. (2008). *Astrobiology*. doi:10.1089/ast.2008.0819

  > The NASA Astrobiology Roadmap provides guidance for research and technology development across the NASA enterprises that encompass the space, Earth, and biological sciences. The ongoing development of astrobiology roadmaps embodies the contributions of diverse scientists and technologists from government, universities, and private institutions. The Roadmap addresses three basic questions: how does life begin and evolve, does life exist elsewhere in the universe, and what is the future of life on Earth and beyond? Seven Science Goals outline the following key domains of investigation: understanding the nature and distribution of habitable environments in the universe, exploring for habitable environments and life in our own Solar System, understanding the emergence of life, determining how early life on Earth interacted and evolved with its changing environment, understanding the evolutionary mechanisms and environmental limits of life, determining the principles that will shape life in the future, and recognizing signatures of life on other worlds and on early Earth. For each of these goals, Science Objectives outline more specific high priority efforts for the next three to five years. These eighteen objectives are being integrated with NASA strategic planning.

- Mayor and Queloz (1995) — the first planet around a Sun-like star.

  A Jupiter-mass companion to a solar-type star. Mayor & Queloz (1995). *Nature*. doi:10.1038/378355a0

B.3.2 Concepts and the neutral theory

- King and Jukes (1969) — non-Darwinian evolution.

  Non-Darwinian Evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift.. King & Jukes (1969). *Science*. doi:10.1126/science.164.3881.788

- Ohta (1973) — the nearly neutral theory.

  Slightly Deleterious Mutant Substitutions in Evolution. Ohta (1973). *Nature*. doi:10.1038/246096a0

- Pauling and Zuckerkandl (1963), Langley and Fitch (1974) — the molecular clock and its constancy.

  Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. Pauling & Zuckerkandl (1963). *Acta Chemica Scandinavica*.

  An examination of the constancy of the rate of molecular evolution. Langley & Fitch (1974). *Journal of Molecular Evolution*.

- Smith and Haigh (1974) — the hitchhiking effect.

  The hitch-hiking effect of a favourable gene. Smith & Haigh (1974). *Genetical Research*. doi:10.1017/s0016672300014634

  > When a selectively favourable gene substitution occurs in a population, changes in gene frequencies will occur at closely linked loci. In the case of a neutral polymorphism, average heterozygosity will be reduced to an extent which varies with distance from the substituted locus. The aggregate effect of substitution on neutral polymorphism is estimated; in populations of total size 10 6 or more (and perhaps of 10 4 or more), this effect will be more important than that of random fixation. This may explain why the extent of polymorphism in natural populations does not vary as much as one would expect from a consideration of the equilibrium between mutation and random fixation in populations of different sizes. For a selectively maintained polymorphism at a linked locus, this process will only be important in the long run if it leads

> to complete fixation. If the selective coefficients at the linked locus are small compared to those at the substituted locus, it is shown that the probability of complete fixation at the linked locus is approximately exp (– Nc), where c is the recombinant fraction and N the population size. It follows that in a large population a selective substitution can occur in a cistron without eliminating a selectively maintained polymorphism in the same cistron.

- Charlesworth (2009) — effective population size and molecular evolution.

  Effective population size and patterns of molecular evolution and variation. Charlesworth (2009). *Nature Reviews Genetics*. doi:10.1038/nrg2526

  > The effective size of a population, N(e), determines the rate of change in the composition of a population caused by genetic drift, which is the random sampling of genetic variants in a finite population. N(e) is crucial in determining the level of variability in a population, and the effectiveness of selection relative to drift. This article reviews the properties of N(e) in a variety of different situations of biological interest, and the factors that influence it. In particular, the action of selection means that N(e) varies across the genome, and advances in genomic techniques are giving new insights into how selection shapes N(e).

- Lynch and Conery (2003), Lynch et al. (2016) — genome complexity and the drift barrier.

  The Origins of Genome Complexity. Lynch & Conery (2003). *Science*. doi:10.1126/science.1089370

  > Complete genomic sequences from diverse phylogenetic lineages reveal notable increases in genome complexity from prokaryotes to multicellular eukaryotes. The changes include gradual increases in gene number, resulting from the retention of duplicate genes, and more abrupt increases in the abundance of spliceosomal introns and mobile genetic elements. We argue that many of these modifications emerged passively in response to the long-term population-size reductions that accompanied increases in organism size. According to this model, much of the restructuring of eukaryotic genomes was initiated by non-adaptive processes, and this in turn provided novel substrates for the secondary evolution of phenotypic complexity by natural selection. The enormous long-term effective population sizes of prokaryotes may impose a substantial barrier to the evolution of complex genomes and morphologies.

  Genetic drift, selection and the evolution of the mutation rate. Lynch et al. (2016). *Nature Reviews Genetics*. doi:10.1038/nrg.2016.104

  > As one of the few cellular traits that can be quantified across the tree of life, DNA-replication fidelity provides an excellent platform for understanding fundamental evolutionary processes. Furthermore, because mutation is the ultimate source of all genetic variation, clarifying why mutation rates vary is crucial for understanding all areas of biology. A potentially revealing hypothesis for mutation-rate evolution is that natural selection primarily operates to improve replication fidelity, with the ultimate limits to what can be achieved set by the power of random genetic drift. This drift-barrier hypothesis is consistent with comparative measures of mutation rates, provides a simple explanation for the existence of error-prone polymerases and yields a formal counter-argument to the view that selection fine-tunes gene-specific mutation rates.

Papers on phylogeny and the tree of life are listed in Chapter 4.

B.3.3 Homology, orthology and duplication

- ⋆ Altenhoff et al. (2019) — inferring orthology and paralogy.

  Inferring Orthology and Paralogy. Altenhoff et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_5

  > The distinction between orthologs and paralogs, genes that started diverging by speciation versus duplication, is relevant in a wide range of contexts, most notably phylogenetic tree inference and protein function annotation. In this chapter, we provide an overview of the methods used to infer orthology and paralogy. We survey both graph-based approaches (and their various grouping strategies) and tree-based approaches, which solve the more general problem of gene/species tree reconciliation. We discuss conceptual differences among the various orthology inference methods and databases and examine the

difficult issue of verifying and benchmarking orthology predictions. Finally, we review typical applications of orthologous genes, groups, and reconciled trees and conclude with thoughts on future methodological developments.

- Fitch (1970), Fitch (2000) — homology defined.

  Distinguishing Homologous from Analogous Proteins. Fitch (1970). *Systematic Zoology*. doi:10.2307/2412448

  Homology. Fitch (2000). *Trends in Genetics*. doi:10.1016/s0168-9525(00)02005-9

  > There are many problems relating to defining the terminology used to describe various biological relationships and getting agreement on which definitions are best. Here, I examine 15 terminological problems, all of which are current, and all of which relate to the usage of homology and its associated terms. I suggest a set of definitions that are intended to be totally consistent among themselves and also as consistent as possible with most current usage.

- Tatusov et al. (1997) — COGs.

  A Genomic Perspective on Protein Families. Tatusov et al. (1997). *Science*. doi:10.1126/science.278.5338.631

  > In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

- Remm et al. (2001) — InParanoid.

  Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Remm et al. (2001). *Journal of Molecular Biology*. doi:10.1006/jmbi.2000.5197

  > Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species. Such genes have often retained identical biological roles in the present-day organisms. It is hence important to identify orthologs for transferring functional information between genes in different organisms with a high degree of reliability. For example, orthologs of human proteins are often functionally characterized in model organisms. Unfortunately, orthology analysis between human and e.g. invertebrates is often complex because of large numbers of paralogs within protein families. Paralogs that predate the species split, which we call out-paralogs, can easily be confused with true orthologs. Paralogs that arose after the species split, which we call in-paralogs, however, are bona fide orthologs by definition. Orthologs and in-paralogs are typically detected with phylogenetic methods, but these are slow and difficult to automate. Automatic clustering methods based on two-way best genome-wide matches on the other hand, have so far not separated in-paralogs from out-paralogs effectively. We present a fully automatic method for finding orthologs and in-paralogs from two species. Ortholog clusters are seeded with a two-way best pairwise match, after which an algorithm for adding in-paralogs is applied. The method bypasses multiple alignments and phylogenetic trees, which can be slow and error-prone steps in classical ortholog detection. Still, it robustly detects complex orthologous relationships and assigns confidence values for both orthologs and in-paralogs. The program, called INPARANOID, was tested on all completely sequenced eukaryotic genomes. To assess the quality of INPARANOID results, ortholog clusters were generated from a dataset of worm and mammalian transmembrane proteins, and were compared to clusters derived by manual tree-based ortholog detection methods. This study led to the identification with a high degree of confidence of over a dozen novel worm-mammalian ortholog assignments that were previously undetected because of shortcomings of phylogenetic methods.A WWW server that allows searching for orthologs between human and several fully sequenced genomes is installed at http://www.cgb.ki.se/inparanoid/. This is the first comprehensive resource with orthologs of all fully sequenced eukaryotic genomes. Programs and tables of orthology assignments are available from the same location.

- Li et al. (2003) — OrthoMCL.

  OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Li et al. (2003). *Genome Research*. doi:10.1101/gr.1224503

  > The identification of orthologous groups is useful for genome annotation, studies on gene/protein evolution, comparative genomics, and the identification of taxonomically restricted sequences. Methods successfully exploited for prokaryotic genome analysis have proved difficult to apply to eukaryotes, however, as larger genomes may contain multiple paralogous genes, and sequence information is often incomplete. OrthoMCL provides a scalable method for constructing orthologous groups across multiple eukaryotic taxa, using a Markov Cluster algorithm to group (putative) orthologs and paralogs. This method performs similarly to the INPARANOID algorithm when applied to two genomes, but can be extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO, but improved recognition of "recent" paralogs permits overlapping EGO groups representing the same gene to be merged. Comparison with previously assigned EC annotations suggests a high degree of reliability, implying utility for automated eukaryotic genome annotation. OrthoMCL has been applied to the proteome data set from seven publicly available genomes (human, fly, worm, yeast, Arabidopsis, the malaria parasite Plasmodium falciparum, and Escherichia coli). A Web interface allows queries based on individual genes or user-defined phylogenetic patterns (http://www.cbil.upenn.edu/gene-family). Analysis of clusters incorporating P. falciparum genes identifies numerous enzymes that were incompletely annotated in first-pass annotation of the parasite genome.

- Huerta-Cepas et al. (2018) — eggNOG.

  eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Huerta-Cepas et al. (2018). *Nucleic Acids Research*. doi:10.1093/nar/gky1085

  > eggNOG is a public database of orthology relationships, gene evolutionary histories and functional annotations. Here, we present version 5.0, featuring a major update of the underlying genome sets, which have been expanded to 4445 representative bacteria and 168 archaea derived from 25 038 genomes, as well as 477 eukaryotic organisms and 2502 viral proteomes that were selected for diversity and filtered by genome quality. In total, 4.4M orthologous groups (OGs) distributed across 379 taxonomic levels were computed together with their associated sequence alignments, phylogenies, HMM models and functional descriptors. Precomputed evolutionary analysis provides fine-grained resolution of duplication/speciation events within each OG. Our benchmarks show that, despite doubling the amount of genomes, the quality of orthology assignments and functional annotations (80% coverage) has persisted without significant changes across this update. Finally, we improved eggNOG online services for fast functional annotation and orthology prediction of custom genomics or metagenomics datasets. All precomputed data are publicly available for downloading or via API queries at http://eggnog.embl.de.

- Altenhoff et al. (2016) — benchmarking orthology inference.

  Standardized benchmarking in the quest for orthologs. Altenhoff et al. (2016). *Nature Methods*. doi:10.1038/nmeth.3830

  > Achieving high accuracy in orthology inference is essential for many comparative, evolutionary and functional genomic analyses, yet the true evolutionary history of genes is generally unknown and orthologs are used for very different applications across phyla, requiring different precision-recall trade-offs. As a result, it is difficult to assess the performance of orthology inference methods. Here, we present a community effort to establish standards and an automated web-based service to facilitate orthology benchmarking. Using this service, we characterize 15 well-established inference methods and resources on a battery of 20 different benchmarks. Standardized benchmarking provides a way for users to identify the most effective methods for the problem at hand, sets a minimum requirement for new tools and resources, and guides the development of more accurate orthology inference methods.

- Nehrt et al. (2011), Altenhoff et al. (2012) — the ortholog conjecture debate.

  Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. Nehrt et al. (2011). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002073

> A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the "ortholog conjecture"). Many methods used to computationally predict protein function are based on this assumption, even though it is largely untested. Here we present the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We use the experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities. Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. We also find that paralogous pairs residing on the same chromosome are more functionally similar than those on different chromosomes, perhaps due to higher levels of interlocus gene conversion between these pairs. In addition to offering implications for the computational prediction of protein function, our results shed light on the relationship between sequence divergence and functional divergence. We conclude that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act.

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. Altenhoff et al. (2012). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002514

> The function of most proteins is not determined experimentally, but is extrapolated from homologs. According to the "ortholog conjecture", or standard model of phylogenomics, protein function changes rapidly after duplication, leading to paralogs with different functions, while orthologs retain the ancestral function. We report here that a comparison of experimentally supported functional annotations among homologs from 13 genomes mostly supports this model. We show that to analyze GO annotation effectively, several confounding factors need to be controlled: authorship bias, variation of GO term frequency among species, variation of background similarity among species pairs, and propagated annotation bias. After controlling for these biases, we observe that orthologs have generally more similar functional annotations than paralogs. This is especially strong for sub-cellular localization. We observe only a weak decrease in functional similarity with increasing sequence divergence. These findings hold over a large diversity of species; notably orthologs from model organisms such as E. coli, yeast or mouse have conserved function with human proteins.

- Ohno et al. (1968) — evolution by gene duplication.

  Evolution from fish to mammals by gene duplication. Ohno et al. (1968). *Hereditas*.

- Force et al. (1999) — subfunctionalisation (the DDC model).

  Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Force et al. (1999). *Genetics*. doi:10.1093/genetics/151.4.1531

  > The origin of organismal complexity is generally thought to be tightly coupled to the evolution of new gene functions arising subsequent to gene duplication. Under the classical model for the evolution of duplicate genes, one member of the duplicated pair usually degenerates within a few million years by accumulating deleterious mutations, while the other duplicate retains the original function. This model further predicts that on rare occasions, one duplicate may acquire a new adaptive function, resulting in the preservation of both members of the pair, one with the new function and the other retaining the old. However, empirical data suggest that a much greater proportion of gene duplicates is preserved than predicted by the classical model. Here we present a new conceptual framework for understanding the evolution of duplicate genes that may help explain this conundrum. Focusing on the regulatory complexity of eukaryotic genes, we show how complementary degenerative mutations in different regulatory elements of duplicated genes can facilitate the preservation of both duplicates, thereby increasing long-term opportunities for the evolution of new gene functions. The duplication-degeneration-complementation (DDC) model predicts that (1) degenerative mutations in regulatory elements can increase rather than reduce the probability of duplicate gene preservation and (2) the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions rather than the evolution of new functions. We present several examples (including analysis of a new engrailed gene in zebrafish) that appear to be consistent with the DDC model, and we suggest several analytical and experimental approaches for determining whether the complementary loss of gene subfunctions or the acquisition of novel functions are likely to be the primary mechanisms for the preservation of gene duplicates. For a newly duplicated paralog, survival depends on the outcome of the race between entropic decay and chance acquisition of an advantageous regulatory mutation. Sidow 1996(p. 717) On one hand, it may fix an advantageous allele giving it a slightly different, and selectable,

> function from its original copy. This initial fixation provides substantial protection against future fixation of null mutations, allowing additional mutations to accumulate that refine functional differentiation. Alternatively, a duplicate locus can instead first fix a null allele, becoming a pseudogene. Walsh 1995 (p. 426) Duplicated genes persist only if mutations create new and essential protein functions, an event that is predicted to occur rarely. Nadeau and Sankoff 1997 (p. 1259) Thus overall, with complex metazoans, the major mechanism for retention of ancient gene duplicates would appear to have been the acquisition of novel expression sites for developmental genes, with its accompanying opportunity for new gene roles underlying the progressive extension of development itself. Cooke et al. 1997 (p. 362)

- Innan and Kondrashov (2010) — models of duplicate gene evolution.

  The evolution of gene duplications: classifying and distinguishing between models. Innan & Kondrashov (2010). *Nature Reviews Genetics*. doi:10.1038/nrg2689

  > Gene duplications and their subsequent divergence play an important part in the evolution of novel gene functions. Several models for the emergence, maintenance and evolution of gene copies have been proposed. However, a clear consensus on how gene duplications are fixed and maintained in genomes is lacking. Here, we present a comprehensive classification of the models that are relevant to all stages of the evolution of gene duplications. Each model predicts a unique combination of evolutionary dynamics and functional properties. Setting out these predictions is an important step towards identifying the main mechanisms that are involved in the evolution of gene duplications.

- Papp et al. (2003) — dosage sensitivity.

  Dosage sensitivity and the evolution of gene families in yeast. Papp et al. (2003). *Nature*. doi:10.1038/nature01771

  > According to what we term the balance hypothesis, an imbalance in the concentration of the subcomponents of a protein-protein complex can be deleterious. If so, there are two consequences: first, both underexpression and overexpression of protein complex subunits should lower fitness, and second, the accuracy of transcriptional co-regulation of subunits should reflect the deleterious consequences of imbalance. Here we show that all these predictions are upheld in yeast (Saccharomyces cerevisiae). This supports the hypothesis that dominance is a by-product of physiology and metabolism rather than the result of selection to mask the deleterious effects of mutations. Beyond this, single-gene duplication of protein subunits is expected to be harmful, as this, too, leads to imbalance. As then expected, we find that members of large gene families are rarely involved in complexes. The balance hypothesis therefore provides a single theoretical framework for understanding components both of dominance and of gene family size.

- Wolfe and Shields (1997), Kellis et al. (2004) — whole-genome duplication in yeast.

  Molecular evidence for an ancient duplication of the entire yeast genome. Wolfe & Shields (1997). *Nature*. doi:10.1038/42711

  > Gene duplication is an important source of evolutionary novelty. Most duplications are of just a single gene, but Ohno proposed that whole-genome duplication (polyploidy) is an important evolutionary mechanism. Many duplicate genes have been found in Saccharomyces cerevisiae, and these often seem to be phenotypically redundant. Here we show that the arrangement of duplicated genes in the S. cerevisiae genome is consistent with Ohno's hypothesis. We propose a model in which this species is a degenerate tetraploid resulting from a whole-genome duplication that occurred after the divergence of Saccharomyces from Kluyveromyces. Only a small fraction of the genes were subsequently retained in duplicate (most were deleted), and gene order was rearranged by many reciprocal translocations between chromosomes. Protein pairs derived from this duplication event make up 13% of all yeast proteins, and include pairs of transcription factors, protein kinases, myosins, cyclins and pheromones. Tetraploidy may have facilitated the evolution of anaerobic fermentation in Saccharomyces.

  Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Kellis et al. (2004). *Nature*. doi:10.1038/nature02424

> Whole-genome duplication followed by massive gene loss and specialization has long been postulated as a powerful mechanism of evolutionary innovation. Recently, it has become possible to test this notion by searching complete genome sequence for signs of ancient duplication. Here, we show that the yeast Saccharomyces cerevisiae arose from ancient whole-genome duplication, by sequencing and analysing Kluyveromyces waltii, a related yeast species that diverged before the duplication. The two genomes are related by a 1:2 mapping, with each region of K. waltii corresponding to two regions of S. cerevisiae, as expected for whole-genome duplication. This resolves the long-standing controversy on the ancestry of the yeast genome, and makes it possible to study the fate of duplicated genes directly. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution, and allowing us to distinguish ancestral and derived functions.

- Dehal and Boore (2005) — two rounds of genome duplication in vertebrates.

  Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. Dehal & Boore (2005). *PLoS Biology*. doi:10.1371/journal.pbio.0030314

  > The hypothesis that the relatively large and complex vertebrate genome was created by two ancient, whole genome duplications has been hotly debated, but remains unresolved. We reconstructed the evolutionary relationships of all gene families from the complete gene sets of a tunicate, fish, mouse, and human, and then determined when each gene duplicated relative to the evolutionary tree of the organisms. We confirmed the results of earlier studies that there remains little signal of these events in numbers of duplicated genes, gene tree topology, or the number of genes per multigene family. However, when we plotted the genomic map positions of only the subset of paralogous genes that were duplicated prior to the fish-tetrapod split, their global physical organization provides unmistakable evidence of two distinct genome duplication events early in vertebrate evolution indicated by clear patterns of four-way paralogous regions covering a large part of the human genome. Our results highlight the potential for these large-scale genomic events to have driven the evolutionary success of the vertebrate lineage.

B.3.4 Protein evolution

- Apic et al. (2001), Björklund et al. (2005) — domain combinations and rearrangements.

  Domain combinations in archaeal, eubacterial and eukaryotic proteomes. Apic et al. (2001). *Journal of Molecular Biology*. doi:10.1006/jmbi.2001.4776

  > There is a limited repertoire of domain families that are duplicated and combined in different ways to form the set of proteins in a genome. Proteins are gene products, and at the level of genes, duplication, recombination, fusion and fission are the processes that produce new genes. We attempt to gain an overview of these processes by studying the evolutionary units in proteins, domains, in the protein sequences of 40 genomes. The domain and superfamily definitions in the Structural Classification of Proteins Database are used, so that we can view all pairs of adjacent domains in genome sequences in terms of their superfamily combinations. We find 783 out of the 859 superfamilies in SCOP in these genomes, and the 783 families occur in 1307 pairwise combinations. Most families are observed in combination with one or two other families, while a few families are very versatile in their combinatorial behaviour; 209 families do not make combinations with other families. This type of pattern can be described as a scale-free network. We also study the N to C-terminal orientation of domain pairs and domain repeats. The phylogenetic distribution of domain combinations is surveyed, to establish the extent of common and kingdom-specific combinations. Of the kingdom-specific combinations, significantly more combinations consist of families present in all three kingdoms than of families present in one or two kingdoms. Hence, we are led to conclude that recombination between common families, as compared to the invention of new families and recombination among these, has also been a major contribution to the evolution of kingdom-specific and species-specific functions in organisms in all three kingdoms. Finally, we compare the set of the domain combinations in the genomes to those in the RCSB Protein Data Bank, and discuss the implications for structural genomics.

  Domain Rearrangements in Protein Evolution. Björklund et al. (2005). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2005.08.067

  > Most eukaryotic proteins are multi-domain proteins that are created from fusions of genes, deletions and internal repetitions. An investigation of such evolutionary events requires a method to find the domain architecture from which each protein originates. Therefore, we defined a novel measure, domain distance,

> which is calculated as the number of domains that differ between two domain architectures. Using this measure the evolutionary events that distinguish a protein from its closest ancestor have been studied and it was found that indels are more common than internal repetition and that the exchange of a domain is rare. Indels and repetitions are common at both the N and C-terminals while they are rare between domains. The evolution of the majority of multi-domain proteins can be explained by the stepwise insertions of single domains, with the exception of repeats that sometimes are duplicated several domains in tandem. We show that domain distances agree with sequence similarity and semantic similarity based on gene ontology annotations. In addition, we demonstrate the use of the domain distance measure to build evolutionary trees. Finally, the evolution of multi-domain proteins is exemplified by a closer study of the evolution of two protein families, non-receptor tyrosine kinases and RhoGEFs.

- Kummerfeld and Teichmann (2005) — fusion and fission.

  Relative rates of gene fusion and fission in multi-domain proteins. Kummerfeld & Teichmann (2005). *Trends in Genetics*. doi:10.1016/j.tig.2004.11.007

  > During evolution genes can produce more complex proteins by gene fusion or less complex proteins by gene fission. Considering proteins from 131 completely sequenced genomes from all three kingdoms of life, we identified 2869 groups of multi-domain proteins as a single protein in certain organisms and as two or more smaller proteins with equivalent domain architectures in other organisms. We found that fusion events are approximately four times more common than fission events, and we established that, in most cases, any particular fusion or fission event only occurred once during the course of evolution.

- Levy et al. (2008) — assembly of complexes reflects evolution.

  Assembly reflects evolution of protein complexes. Levy et al. (2008). *Nature*. doi:10.1038/nature06942

  > A homomer is formed by self-interacting copies of a protein unit. This is functionally important, as in allostery, and structurally crucial because mis-assembly of homomers is implicated in disease. Homomers are widespread, with 50-70% of proteins with a known quaternary state assembling into such structures. Despite their prevalence, their role in the evolution of cellular machinery and the potential for their use in the design of new molecular machines, little is known about the mechanisms that drive formation of homomers at the level of evolution and assembly in the cell. Here we present an analysis of over 5,000 unique atomic structures and show that the quaternary structure of homomers is conserved in over 70% of protein pairs sharing as little as 30% sequence identity. Where quaternary structure is not conserved among the members of a protein family, a detailed investigation revealed well-defined evolutionary pathways by which proteins transit between different quaternary structure types. Furthermore, we show by perturbing subunit interfaces within complexes and by mass spectrometry analysis, that the (dis)assembly pathway mimics the evolutionary pathway. These data represent a molecular analogy to Haeckel's evolutionary paradigm of embryonic development, where an intermediate in the assembly of a complex represents a form that appeared in its own evolutionary history. Our model of self-assembly allows reliable prediction of evolution and assembly of a complex solely from its crystal structure.

- Björklund et al. (2006) — expansion of domain repeats.

  Expansion of Protein Domain Repeats. Björklund et al. (2006). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.0020114

  > Many proteins, especially in eukaryotes, contain tandem repeats of several domains from the same family. These repeats have a variety of binding properties and are involved in protein-protein interactions as well as binding to other ligands such as DNA and RNA. The rapid expansion of protein domain repeats is assumed to have evolved through internal tandem duplications. However, the exact mechanisms behind these tandem duplications are not well-understood. Here, we have studied the evolution, function, protein structure, gene structure, and phylogenetic distribution of domain repeats. For this purpose we have assigned Pfam-A domain families to 24 proteomes with more sensitive domain assignments in the repeat regions. These assignments confirmed previous findings that eukaryotes, and in particular vertebrates, contain a much higher fraction of proteins with repeats compared with prokaryotes. The internal sequence similarity in each protein revealed that the domain repeats are often expanded through duplications of several domains at a time, while the duplication of one domain is less common. Many of the repeats appear to have been duplicated in the middle of the repeat region. This is in strong contrast to the evolution of other proteins that mainly works through additions of single domains at either terminus. Further, we found that some domain families show distinct duplication patterns, e.g., nebulin domains

have mainly been expanded with a unit of seven domains at a time, while duplications of other domain families involve varying numbers of domains. Finally, no common mechanism for the expansion of all repeats could be detected. We found that the duplication patterns show no dependence on the size of the domains. Further, repeat expansion in some families can possibly be explained by shuffling of exons. However, exon shuffling could not have created all repeats.

- Kajava (2012) — tandem repeats in proteins.

  Tandem repeats in proteins: From sequence to structure. Kajava (2012). *Journal of Structural Biology*. doi:10.1016/j.jsb.2011.08.009

  > The bioinformatics analysis of proteins containing tandem repeats requires special computer programs and databases, since the conventional approaches predominantly developed for globular domains have limited success. Here, I survey bioinformatics tools which have been developed recently for identification and proteome-wide analysis of protein repeats. The last few years have also been marked by an emergence of new 3D structures of these proteins. Appraisal of the known structures and their classification uncovers a straightforward relationship between their architecture and the length of the repetitive units. This relationship and the repetitive character of structural folds suggest rules for better prediction of the 3D structures of such proteins. Furthermore, bioinformatics approaches combined with low resolution structural data, from biophysical techniques, especially, the recently emerged cryo-electron microscopy, lead to reliable prediction of the protein repeat structures and their mode of binding with partners within molecular complexes. This hybrid approach can actively be used for structural and functional annotations of proteomes.

- Brown et al. (2011), Lee et al. (2014), Light et al. (2013) — evolution of disordered regions.

  Evolution and disorder. Brown et al. (2011). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2011.02.005

  > The evolution of disordered proteins or regions of proteins differs from that of ordered proteins because of the differences in their sequence composition, intramolecular contacts, and function. Recent assessments of disordered protein evolution at the sequence, structural, and functional levels support this hypothesis. Disordered proteins have a different pattern of accepted point mutations, exhibit higher rates of insertions and deletions, and generally, but not always, evolve more rapidly than ordered proteins. Even with these high rates of sequence evolution, a few examples have shown that disordered proteins maintain their flexibility under physiological conditions, and it is hypothesized that they maintain specific structural ensembles.

  Classification of Intrinsically Disordered Regions and Proteins. van der Lee et al. (2014). *Chemical Reviews*. doi:10.1021/cr400525m

  > Characterization of unannotated and uncharacterized protein segments is expected to lead to the discovery of novel functions as well as provide important insights into existing biological processes. In addition, it is likely to shed new light on molecular mechanisms of diseases that are not yet fully understood. The classical concept implies that protein sequence defines structure, which in turn determines function; that is, function can be inferred from the sequence and its structure. Even when protein sequences diverge during evolution, for example, after gene duplication, the overall fold of their structures remains roughly the same. Therefore, structural similarity between proteins can reveal distant evolutionary relationships that are not easily detectable using sequence-based methods.

  Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. Light et al. (2013). *Molecular Biology and Evolution*. doi:10.1093/molbev/mst157

  > Proteins evolve not only through point mutations but also by insertion and deletion events, which affect the length of the protein. It is well known that such indel events most frequently occur in surface-exposed loops. However, detailed analysis of indel events in distantly related and fast-evolving proteins is hampered by the difficulty involved in correctly aligning such sequences. Here, we circumvent this problem by first only analyzing homologous proteins based on length variation rather than pairwise alignments. Using this approach, we find a surprisingly strong relationship between difference in length and difference in the number of intrinsically disordered residues, where up to three quarters of the length variation can be explained by changes in the number of intrinsically disordered residues. Further, we find that disorder is common in both insertions and deletions. A more detailed analysis reveals that indel

events do not induce disorder but rather that already disordered regions accrue indels, suggesting that there is a lowered selective pressure for indels to occur within intrinsically disordered regions.

- Göbel et al. (1994), Weigt et al. (2009), Marks et al. (2011) — correlated mutations and coevolution.

Correlated mutations and residue contacts in proteins. Göbel et al. (1994). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.340180402

> The maintenance of protein function and structure constrains the evolution of amino acid sequences. This fact can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions. Here we present a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. We then use these correlations to predict contact maps for each of 11 protein families and compare the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranges from 37 to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. Predicted contact maps can be used as input for the calculation of protein tertiary structure, either from sequence information alone or in combination with experimental information.

Identification of direct residue contacts in protein–protein interaction by message passing. Weigt et al. (2009). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0805923106

> Understanding the molecular determinants of specificity in protein-protein interaction is an outstanding challenge of postgenome biology. The availability of large protein databases generated from sequences of hundreds of bacterial genomes enables various statistical approaches to this problem. In this context covariance-based methods have been used to identify correlation between amino acid positions in interacting proteins. However, these methods have an important shortcoming, in that they cannot distinguish between directly and indirectly correlated residues. We developed a method that combines covariance analysis with global inference analysis, adopted from use in statistical physics. Applied to a set of >2,500 representatives of the bacterial two-component signal transduction system, the combination of covariance with global inference successfully and robustly identified residue pairs that are proximal in space without resorting to ad hoc tuning parameters, both for heterointeractions between sensor kinase (SK) and response regulator (RR) proteins and for homointeractions between RR proteins. The spectacular success of this approach illustrates the effectiveness of the global inference approach in identifying direct interaction based on sequence information alone. We expect this method to be applicable soon to interaction surfaces between proteins present in only 1 copy per genome as the number of sequenced genomes continues to expand. Use of this method could significantly increase the potential targets for therapeutic intervention, shed light on the mechanism of protein-protein interaction, and establish the foundation for the accurate prediction of interacting protein partners.

Protein 3D Structure Computed from Evolutionary Sequence Variation. Marks et al. (2011). *PLoS ONE*. doi:10.1371/journal.pone.0028766

> The evolutionary trajectory of a protein through sequence space is constrained by its function. Collections of sequence homologs record the outcomes of millions of evolutionary experiments in which the protein evolves according to these constraints. Deciphering the evolutionary record held in these sequences and exploiting it for predictive and engineering purposes presents a formidable challenge. The potential benefit of solving this challenge is amplified by the advent of inexpensive high-throughput genomic sequencing.In this paper we ask whether we can infer evolutionary constraints from a set of sequence homologs of a protein. The challenge is to distinguish true co-evolution couplings from the noisy set of observed correlations. We address this challenge using a maximum entropy model of the protein sequence, constrained by the statistics of the multiple sequence alignment, to infer residue pair couplings. Surprisingly, we find that the strength of these inferred couplings is an excellent predictor of residue-residue proximity in folded structures. Indeed, the top-scoring residue couplings are sufficiently accurate and well-distributed to define the 3D protein fold with remarkable accuracy.We quantify this observation by computing, from sequence alone, all-atom 3D structures of fifteen test proteins from different fold classes, ranging in size from 50 to 260 residues, including a G-protein coupled receptor. These blinded inferences are de novo, i.e., they do not use homology modeling or sequence-similar fragments from known structures. The co-evolution signals provide sufficient information to determine accurate 3D protein structure to 2.7-4.8 Å C(α)-RMSD error relative to the observed structure, over at least two-thirds of the protein (method called EVfold, details at http://EVfold.org). This discovery provides insight into essential interactions constraining protein evolution and will facilitate a comprehensive survey of

the universe of protein structures, new strategies in protein and drug design, and the identification of functional genetic variants in normal and disease genomes.

- Barbosa-Morais et al. (2012), Merkin et al. (2012), Tress et al. (2017) — evolution and function of alternative splicing.

  The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Barbosa-Morais et al. (2012). *Science*. doi:10.1126/science.1230612

  > How species with similar repertoires of protein-coding genes differ so markedly at the phenotypic level is poorly understood. By comparing organ transcriptomes from vertebrate species spanning ~350 million years of evolution, we observed significant differences in alternative splicing complexity between vertebrate lineages, with the highest complexity in primates. Within 6 million years, the splicing profiles of physiologically equivalent organs diverged such that they are more strongly related to the identity of a species than they are to organ type. Most vertebrate species-specific splicing patterns are cis-directed. However, a subset of pronounced splicing changes are predicted to remodel protein interactions involving trans-acting regulators. These events likely further contributed to the diversification of splicing and other transcriptomic changes that underlie phenotypic differences among vertebrate species.

  Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Merkin et al. (2012). *Science*. doi:10.1126/science.1228186

  > Most mammalian genes produce multiple distinct messenger RNAs through alternative splicing, but the extent of splicing conservation is not clear. To assess tissue-specific transcriptome variation across mammals, we sequenced complementary DNA from nine tissues from four mammals and one bird in biological triplicate, at unprecedented depth. We find that while tissue-specific gene expression programs are largely conserved, alternative splicing is well conserved in only a subset of tissues and is frequently lineage-specific. Thousands of previously unknown, lineage-specific, and conserved alternative exons were identified; widely conserved alternative exons had signatures of binding by MBNL, PTB, RBFOX, STAR, and TIA family splicing factors, implicating them as ancestral mammalian splicing regulators. Our data also indicate that alternative splicing often alters protein phosphorylatability, delimiting the scope of kinase signaling.

  Alternative Splicing May Not Be the Key to Proteome Complexity. Tress et al. (2017). *Trends in Biochemical Sciences*. doi:10.1016/j.tibs.2016.08.008

  > Alternative splicing is commonly believed to be a major source of cellular protein diversity. However, although many thousands of alternatively spliced transcripts are routinely detected in RNA-seq studies, reliable large-scale mass spectrometry-based proteomics analyses identify only a small fraction of annotated alternative isoforms. The clearest finding from proteomics experiments is that most human genes have a single main protein isoform, while those alternative isoforms that are identified tend to be the most biologically plausible: those with the most cross-species conservation and those that do not compromise functional domains. Indeed, most alternative exons do not seem to be under selective pressure, suggesting that a large majority of predicted alternative transcripts may not even be translated into proteins.

- Drummond et al. (2005) — why highly expressed proteins evolve slowly.

  Why highly expressed proteins evolve slowly. Drummond et al. (2005). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0504070102

  > Much recent work has explored molecular and population-genetic constraints on the rate of protein sequence evolution. The best predictor of evolutionary rate is expression level, for reasons that have remained unexplained. Here, we hypothesize that selection to reduce the burden of protein misfolding will favor protein sequences with increased robustness to translational missense errors. Pressure for translational robustness increases with expression level and constrains sequence evolution. Using several sequenced yeast genomes, global expression and protein abundance data, and sets of paralogs traceable to an ancient whole-genome duplication in yeast, we rule out several confounding effects and show that expression level explains roughly half the variation in Saccharomyces cerevisiae protein evolutionary rates. We examine causes for expression's dominant role and find that genome-wide tests favor the translational robustness explanation over existing hypotheses that invoke constraints on function or translational efficiency. Our results suggest that proteins evolve at rates largely unrelated to their functions and can explain why highly expressed proteins evolve slowly across the tree of life.

- Taverna and Goldstein (2002) — robustness and marginal stability.

  Why are proteins so robust to site mutations?. Taverna & Goldstein (2002). *Journal of Molecular Biology*. doi:10.1006/jmbi.2001.5226

  > There have been repeated observations that proteins are surprisingly robust to site mutations, enduring significant numbers of substitutions with little change in structure, stability, or function. These results are almost paradoxical in light of what is known about random heteropolymers and the sensitivity of their properties to seemingly trivial mutations. To address this discrepancy, the preservation of biological protein properties in the presence of mutation has been interpreted as indicating the independence of selective pressure on such properties. Such results also lead to the prediction that de novo protein design should be relatively easy, in contrast to what is observed. Here, we use a computational model with lattice proteins to demonstrate how this robustness can result from population dynamics during the evolutionary process. As a result, sequence plasticity may be a characteristic of evolutionarily derived proteins and not necessarily a property of designed proteins. This suggests that this robustness must be re-interpreted in evolutionary terms, and has consequences for our understanding of both in vivo and in vitro protein evolution.

- Doolittle (2010) — the roots of bioinformatics in protein evolution.

  The Roots of Bioinformatics in Protein Evolution. Doolittle (2010). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000875

B.3.5 Selection

- Yang and Bielawski (2000) — statistical methods for detecting molecular adaptation.

  Statistical methods for detecting molecular adaptation. Yang & Bielawski (2000). *Trends in Ecology & Evolution*. doi:10.1016/s0169-5347(00)01994-7

  > The past few years have seen the development of powerful statistical methods for detecting adaptive molecular evolution. These methods compare synonymous and nonsynonymous substitution rates in protein-coding genes, and regard a nonsynonymous rate elevated above the synonymous rate as evidence for darwinian selection. Numerous cases of molecular adaptation are being identified in various systems from viruses to humans. Although previous analyses averaging rates over sites and time have little power, recent methods designed to detect positive selection at individual sites and lineages have been successful. Here, we summarize recent statistical methods for detecting molecular adaptation, and discuss their limitations and possible improvements.

- Nei and Gojobori (1986), Goldman and Yang (1994) — counting methods and codon models.

  Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Nei & Gojobori (1986). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040410

  > Two simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions are presented. Although they give no weights to different types of codon substitutions, these methods give essentially the same results as those obtained by Miyata and Yasunaga's and by Li et al.'s methods. Computer simulation indicates that estimates of synonymous substitutions obtained by the two methods are quite accurate unless the number of nucleotide substitutions per site is very large. It is shown that all available methods tend to give an underestimate of the number of nonsynonymous substitutions when the number is large.

  A codon-based model of nucleotide substitution for protein-coding DNA sequences. Goldman & Yang (1994). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040153

A codon-based model for the evolution of protein-coding DNA sequences is presented for use in phylogenetic estimation. A Markov process is used to describe substitutions between codons. Transition/transversion rate bias and codon usage bias are allowed in the model, and selective restraints at the protein level are accommodated using physicochemical distances between the amino acids coded for by the codons. Analyses of two data sets suggest that the new codon-based model can provide a better fit to data than can nucleotide-based models and can produce more reliable estimates of certain biologically important measures such as the transition/transversion rate ratio and the synonymous/nonsynonymous substitution rate ratio.

- Nielsen and Yang (1998), Zhang (2005) — site and branch-site models.

Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. Nielsen & Yang (1998). *Genetics*. doi:10.1093/genetics/148.3.929

Several codon-based models for the evolution of protein-coding DNA sequences are developed that account for varying selection intensity among amino acid sites. The "neutral model" assumes two categories of sites at which amino acid replacements are either neutral or deleterious. The "positive-selection model" assumes an additional category of positively selected sites at which nonsynonymous substitutions occur at a higher rate than synonymous ones. This model is also used to identify target sites for positive selection. The models are applied to a data set of the V3 region of the HIV-1 envelope gene, sequenced at different years after the infection of one patient. The results provide strong support for variable selection intensity among amino acid sites The neutral model is rejected in favor of the positive-selection model, indicating the operation of positive selection in the region. Positively selected sites are found in both the V3 region and the flanking regions.

Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Zhang (2005). *Molecular Biology and Evolution*. doi:10.1093/molbev/msi237

Detecting positive Darwinian selection at the DNA sequence level has been a subject of considerable interest. However, positive selection is difficult to detect because it often operates episodically on a few amino acid sites, and the signal may be masked by negative selection. Several methods have been developed to test positive selection that acts on given branches (branch methods) or on a subset of sites (site methods). Recently, Yang, Z., and R. Nielsen (2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908-917) developed likelihood ratio tests (LRTs) based on branch-site models to detect positive selection that affects a small number of sites along prespecified lineages. However, computer simulations suggested that the tests were sensitive to the model assumptions and were unable to distinguish between relaxation of selective constraint and positive selection (Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. Mol. Biol. Evol. 21:1332-1339). Here, we describe a modified branch-site model and use it to construct two LRTs, called branch-site tests 1 and 2. We applied the new tests to reanalyze several real data sets and used computer simulation to examine the performance of the two tests by examining their false-positive rate, power, and robustness. We found that test 1 was unable to distinguish relaxed constraint from positive selection affecting the lineages of interest, while test 2 had acceptable false-positive rates and appeared robust against violations of model assumptions. As test 2 is a direct test of positive selection on the lineages of interest, it is referred to as the branch-site test of positive selection and is recommended for use in real data analysis. The test appeared conservative overall, but exhibited better power in detecting positive selection than the branch-based test. Bayes empirical Bayes identification of amino acid sites under positive selection along the foreground branches was found to be reliable, but lacked power.

- Yang (2007), Pond et al. (2004) — PAML and HyPhy.

PAML 4: Phylogenetic Analysis by Maximum Likelihood. Yang (2007). *Molecular Biology and Evolution*. doi:10.1093/molbev/msm088

PAML, currently in version 4, is a package of programs for phylogenetic analyses of DNA and protein sequences using maximum likelihood (ML). The programs may be used to compare and test phylogenetic trees, but their main strengths lie in the rich repertoire of evolutionary models implemented, which can be used to estimate parameters in models of sequence evolution and to test interesting biological hypotheses. Uses of the programs include estimation of synonymous and nonsynonymous rates (d(N) and d(S)) between two protein-coding DNA sequences, inference of positive Darwinian selection through phylogenetic comparison of protein-coding genes, reconstruction of ancestral genes and proteins for

molecular restoration studies of extinct life forms, combined analysis of heterogeneous data sets from multiple gene loci, and estimation of species divergence times incorporating uncertainties in fossil calibrations. This note discusses some of the major applications of the package, which includes example data sets to demonstrate their use. The package is written in ANSI C, and runs under Windows, Mac OSX, and UNIX systems. It is available at -- (http://abacus.gene.ucl.ac.uk/software/paml.html).

HyPhy: hypothesis testing using phylogenies. Pond et al. (2004). *Bioinformatics*. doi:10.1093/bioinformatics/bti079

> Unlabelled The HyPhypackage is designed to provide a flexible and unified platform for carrying out likelihood-based analyses on multiple alignments of molecular sequence data, with the emphasis on studies of rates and patterns of sequence evolution. Availability http://www.hyphy.org Contact muse@stat.ncsu.edu Supplementary information HyPhydocumentation and tutorials are available at http://www.hyphy.org.

- McDonald and Kreitman (1991) — the McDonald–Kreitman test.

Adaptive protein evolution at the Adh locus in Drosophila. McDonald & Kreitman (1991). *Nature*. doi:10.1038/351652a0

> Proteins often differ in amino-acid sequence across species. This difference has evolved by the accumulation of neutral mutations by random drift, the fixation of adaptive mutations by selection, or a mixture of the two. Here we propose a simple statistical test of the neutral protein evolution hypothesis based on a comparison of the number of amino-acid replacement substitutions to synonymous substitutions in the coding region of a locus. If the observed substitutions are neutral, the ratio of replacement to synonymous fixed differences between species should be the same as the ratio of replacement to synonymous polymorphisms within species. DNA sequence data on the Adh locus (encoding alcohol dehydrogenase, EC 1.1.1.1) in three species in the Drosophila melanogaster species subgroup do not fit this expectation; instead, there are more fixed replacement differences between species than expected. We suggest that these excess replacement substitutions result from adaptive fixation of selectively advantageous mutations.

- Kryazhimskiy and Plotkin (2008) — the population genetics of dN/dS.

The Population Genetics of dN/dS. Kryazhimskiy & Plotkin (2008). *PLoS Genetics*. doi:10.1371/journal.pgen.1000304

> Evolutionary pressures on proteins are often quantified by the ratio of substitution rates at nonsynonymous and synonymous sites. The dN/dS ratio was originally developed for application to distantly diverged sequences, the differences among which represent substitutions that have fixed along independent lineages. Nevertheless, the dN/dS measure is often applied to sequences sampled from a single population, the differences among which represent segregating polymorphisms. Here, we study the expected dN/dS ratio for samples drawn from a single population under selection, and we find that in this context, dN/dS is relatively insensitive to the selection coefficient. Moreover, the hallmark signature of positive selection over divergent lineages, dN/dS>1, is violated within a population. For population samples, the relationship between selection and dN/dS does not follow a monotonic function, and so it may be impossible to infer selection pressures from dN/dS. These results have significant implications for the interpretation of dN/dS measurements among population-genetic samples.

B.3.6 Genome evolution

- Ochman et al. (2000), Koonin et al. (2001) — horizontal gene transfer.

Lateral gene transfer and the nature of bacterial innovation. Ochman et al. (2000). *Nature*. doi:10.1038/35012500

> Unlike eukaryotes, which evolve principally through the modification of existing genetic information, bacteria have obtained a significant proportion of their genetic diversity through the acquisition of sequences from distantly related organisms. Horizontal gene transfer produces extremely dynamic genomes in which substantial amounts of DNA are introduced into and deleted from the chromosome. These lateral transfers have effectively changed the ecological and pathogenic character of bacterial species.

Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. Koonin et al. (2001). *Annual Review of Microbiology*. doi:10.1146/annurev.micro.55.1.709

> Comparative analysis of bacterial, archaeal, and eukaryotic genomes indicates that a significant fraction of the genes in the prokaryotic genomes have been subject to horizontal transfer. In some cases, the amount and source of horizontal gene transfer can be linked to an organism's lifestyle. For example, bacterial hyperthermophiles seem to have exchanged genes with archaea to a greater extent than other bacteria, whereas transfer of certain classes of eukaryotic genes is most common in parasitic and symbiotic bacteria. Horizontal transfer events can be classified into distinct categories of acquisition of new genes, acquisition of paralogs of existing genes, and xenologous gene displacement whereby a gene is displaced by a horizontally transferred ortholog from another lineage (xenolog). Each of these types of horizontal gene transfer is common among prokaryotes, but their relative contributions differ in different lineages. The fixation and long-term persistence of horizontally transferred genes suggests that they confer a selective advantage on the recipient organism. In most cases, the nature of this advantage remains unclear, but detailed examination of several cases of acquisition of eukaryotic genes by bacteria seems to reveal the evolutionary forces involved. Examples include isoleucyl-tRNA synthetases whose acquisition from eukaryotes by several bacteria is linked to antibiotic resistance, ATP/ADP translocases acquired by intracellular parasitic bacteria, Chlamydia and Rickettsia, apparently from plants, and proteases that may be implicated in chlamydial pathogenesis.

- W. F. Doolittle (1999) — the tree of life questioned.

Phylogenetic Classification and the Universal Tree. Doolittle (1999). *Science*. doi:10.1126/science.284.5423.2124

> From comparative analyses of the nucleotide sequences of genes encoding ribosomal RNAs and several proteins, molecular phylogeneticists have constructed a ”universal tree of life,” taking it as the basis for a ”natural” hierarchical classification of all living things. Although confidence in some of the tree's early branches has recently been shaken, new approaches could still resolve many methodological uncertainties. More challenging is evidence that most archaeal and bacterial genomes (and the inferred ancestral eukaryotic nuclear genome) contain genes from multiple sources. If ”chimerism” or ”lateral gene transfer” cannot be dismissed as trivial in extent or limited to special categories of genes, then no hierarchical universal classification can be taken as natural. Molecular phylogeneticists will have failed to find the ”true tree,” not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree. However, taxonomies based on molecular sequences will remain indispensable, and understanding of the evolutionary process will ultimately be enriched, not impoverished.

- Tettelin et al. (2005) — the pan-genome.

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial “pan-genome”. Tettelin et al. (2005). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0506758102

> The development of efficient and inexpensive genome sequencing methods has revolutionized the study of human bacterial pathogens and improved vaccine design. Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets. We have generated the genomic sequence of six strains representing the five major disease-causing serotypes of Streptococcus agalactiae, the main cause of neonatal infection in humans. Analysis of these genomes and those available in databases showed that the S. agalactiae species can be described by a pan-genome consisting of a core genome shared by all isolates, accounting for approximately 80% of any single genome, plus a dispensable genome consisting of partially shared and strain-specific genes. Mathematical extrapolation of the data suggests that the gene reservoir available for inclusion in the S. agalactiae pan-genome is vast and that unique genes will continue to be identified even after sequencing hundreds of genomes.

- Huynen and Nimwegen (1998), Qian et al. (2001) — power-law distributions of family sizes.

The frequency distribution of gene family sizes in complete genomes. Huynen & van Nimwegen (1998). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a025959

> We compare the frequency distribution of gene family sizes in the complete genomes of six bacteria (Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Mycoplasma pneumoniae, and Synechocystis sp. PCC6803), two Archaea (Methanococcus jannaschii and Methanobacterium thermoautotrophicum), one eukaryote (Saccharomyces cerevisiae), the vaccinia virus, and the bacteriophage T4. The sizes of the gene families versus their frequencies show power-law distributions that tend to become flatter (have a larger exponent) as the number of genes in the genome increases. Power-law distributions generally occur as the limit distribution of a multiplicative stochastic process with a boundary constraint. We discuss various models that can account for a multiplicative process determining the sizes of gene families in the genome. In particular, we argue that, in order to explain the observed distributions, gene families have to behave in a coherent fashion within the genome; i.e., the probabilities of duplications of genes within a gene family are not independent of each other. Likewise, the probabilities of deletions of genes within a gene family are not independent of each other.

Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. Qian et al. (2001). *Journal of Molecular Biology*. doi:10.1006/jmbi.2001.5079

> Global surveys of genomes measure the usage of essential molecular parts, defined here as protein families, superfamilies or folds, in different organisms. Based on surveys of the first 20 completely sequenced genomes, we observe that the occurrence of these parts follows a power-law distribution. That is, the number of distinct parts (F) with a given genomic occurrence (V) decays as F=aV(-b), with a few parts occurring many times and most occurring infrequently. For a given organism, the distributions of families, superfamilies and folds are nearly identical, and this is reflected in the size of the decay exponent b. Moreover, the exponent varies between different organisms, with those of smaller genomes displaying a steeper decay (i.e. larger b). Clearly, the power law indicates a preference to duplicate genes that encode for molecular parts which are already common. Here, we present a minimal, but biologically meaningful model that accurately describes the observed power law. Although the model performs equally well for all three protein classes, we focus on the occurrence of folds in preference to families and superfamilies. This is because folds are comparatively insensitive to the effects of point mutations that can cause a family member to diverge beyond detectable similarity. In the model, genomes evolve through two basic operations: (i) duplication of existing genes; (ii) net flow of new genes. The flow term is closely related to the exponent b and can accommodate considerable gene loss; however, we demonstrate that the observed data is reproduced best with a net inflow, i.e. with more gene gain than loss. Moreover, we show that prokaryotes have much higher rates of gene acquisition than eukaryotes, probably reflecting lateral transfer. A further natural outcome from our model is an estimation of the fold composition of the initial genome, which potentially relates to the common ancestor for modern organisms. Supplementary material pertaining to this work is available from www.partslist.org/powerlaw.

- Koonin et al. (2002), Nimwegen (2003) — models and scaling laws.

  The structure of the protein universe and genome evolution. Koonin et al. (2002). *Nature*. doi:10.1038/nature01256

  > Despite the practically unlimited number of possible protein sequences, the number of basic shapes in which proteins fold seems not only to be finite, but also to be relatively small, with probably no more than 10,000 folds in existence. Moreover, the distribution of proteins among these folds is highly non-homogeneous -- some folds and superfamilies are extremely abundant, but most are rare. Protein folds and families encoded in diverse genomes show similar size distributions with notable mathematical properties, which also extend to the number of connections between domains in multidomain proteins. All these distributions follow asymptotic power laws, such as have been identified in a wide variety of biological and physical systems, and which are typically associated with scale-free networks. These findings suggest that genome evolution is driven by extremely general mechanisms based on the preferential attachment principle.

  Scaling laws in the functional content of genomes. van Nimwegen (2003). *Trends in Genetics*. doi:10.1016/s0168-9525(03)00203-8

- Koonin (2011) — are there laws of genome evolution?

  Are There Laws of Genome Evolution?. Koonin (2011). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002173

> Research in quantitative evolutionary genomics and systems biology led to the discovery of several universal regularities connecting genomic and molecular phenomic variables. These universals include the log-normal distribution of the evolutionary rates of orthologous genes; the power law-like distributions of paralogous family size and node degree in various biological networks; the negative correlation between a gene's sequence evolution rate and expression level; and differential scaling of functional classes of genes with genome size. The universals of genome evolution can be accounted for by simple mathematical models similar to those used in statistical physics, such as the birth-death-innovation model. These models do not explicitly incorporate selection; therefore, the observed universal regularities do not appear to be shaped by selection but rather are emergent properties of gene ensembles. Although a complete physical theory of evolutionary biology is inconceivable, the universals of genome evolution might qualify as "laws of evolutionary genomics" in the same sense "law" is understood in modern physics.

B.4 Phylogeny

Further papers for Chapter 4.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.4.1 Reviews and books

- Yang and Rannala (2012) — molecular phylogenetics: principles and practice.

  Molecular phylogenetics: principles and practice. Yang & Rannala (2012). *Nature Reviews Genetics*. doi:10.1038/nrg3186

  > Phylogenies are important for addressing various biological questions such as relationships among species or genes, the origin and spread of viral infection and the demographic changes and migration patterns of species. The advancement of sequencing technologies has taken phylogenetic analysis to a new height. Phylogenies have permeated nearly every branch of biology, and the plethora of phylogenetic methods and software packages that are now available may seem daunting to an experimental biologist. Here, we review the major methods of phylogenetic analysis, including parsimony, distance, likelihood and Bayesian methods. We discuss their strengths and weaknesses and provide guidance for their use.

- Felsenstein (2004) — *Inferring Phylogenies*, the standard textbook, with much of the history.

  Inferring Phylogenies. Felsenstein (2004). *Sinauer Associates*.

- Delsuc et al. (2005) — phylogenomics and the reconstruction of the tree of life.

  Phylogenomics and the reconstruction of the tree of life. Delsuc et al. (2005). *Nature Reviews Genetics*. doi:10.1038/nrg1603

  > As more complete genomes are sequenced, phylogenetic analysis is entering a new era - that of phylogenomics. One branch of this expanding field aims to reconstruct the evolutionary history of organisms on the basis of the analysis of their genomes. Recent studies have demonstrated the power of this approach, which has the potential to provide answers to several fundamental evolutionary questions. However, challenges for the future have also been revealed. The very nature of the evolutionary history of organisms and the limitations of current phylogenetic reconstruction methods mean that part of the tree of life might prove difficult, if not impossible, to resolve with confidence.

- Steel and Penny (2000) — parsimony, likelihood and the role of models.

  Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Steel & Penny (2000). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a026364

> Methods such as maximum parsimony (MP) are frequently criticized as being statistically unsound and not being based on any "model." On the other hand, advocates of MP claim that maximum likelihood (ML) has some fundamental problems. Here, we explore the connection between the different versions of MP and ML methods, particularly in light of recent theoretical results. We describe links between the two methods--for example, we describe how MP can be regarded as an ML method when there is no common mechanism between sites (such as might occur with morphological data and certain forms of molecular data). In the process, we clarify certain historical points of disagreement between proponents of the two methodologies, including a discussion of several forms of the ML optimality criterion. We also describe some additional results that shed light on how much needs to be assumed about underlying models of sequence evolution in order to successfully reconstruct evolutionary trees.

- Steenwyk et al. (2023) — incongruence in the phylogenomics era.

  Incongruence in the phylogenomics era. Steenwyk et al. (2023). *Nature Reviews Genetics*. doi:10.1038/s41576-023-00620-x

  > Genome-scale data and the development of novel statistical phylogenetic approaches have greatly aided the reconstruction of a broad sketch of the tree of life and resolved many of its branches. However, incongruence - the inference of conflicting evolutionary histories - remains pervasive in phylogenomic data, hampering our ability to reconstruct and interpret the tree of life. Biological factors, such as incomplete lineage sorting, horizontal gene transfer, hybridization, introgression, recombination and convergent molecular evolution, can lead to gene phylogenies that differ from the species tree. In addition, analytical factors, including stochastic, systematic and treatment errors, can drive incongruence. Here, we review these factors, discuss methodological advances to identify and handle incongruence, and highlight avenues for future research.

B.4.2 History and classic methods

- Hennig (1965) — phylogenetic systematics (cladistics).

  Phylogenetic Systematics. Hennig (1965). *Annual Review of Entomology*. doi:10.1146/annurev.en.10.010165.000525

- Fitch and Margoliash (1967) — one of the first molecular trees.

  Construction of Phylogenetic Trees: A method based on mutation distances as estimated from cytochrome c sequences is of general applicability.. Fitch & Margoliash (1967). *Science*. doi:10.1126/science.155.3760.279

- Fitch (1971) — minimum change on a given topology (parsimony).

  Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Fitch (1971). *Systematic Zoology*. doi:10.2307/2412116

- Sankoff (1975) — weighted parsimony.

  Minimal Mutation Trees of Sequences. Sankoff (1975). *SIAM Journal on Applied Mathematics*. doi:10.1137/0128004

  > Given a finite tree, some of whose vertices are identified with given finite sequences, we show how to construct sequences for all the remaining vertices simultaneously, so as to minimize the total edge-length of the tree. Edge-length is calculated by a metric whose biological significance is the mutational distance between two sequences.

- Felsenstein (1978) — when parsimony is positively misleading: long-branch attraction.

  Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Felsenstein (1978). *Systematic Biology*. doi:10.1093/sysbio/27.4.401

> For some simple three- and four-species cases involving a character with two states, it is determined under what conditions several methods of phylogenetic inference will fail to converge to the true phylogeny as more and more data are accumulated. The methods are the Camin-Sokal parsimony method, the compatibility method, and Farris's unrooted Wagner tree parsimony method. In all cases the conditions for this failure (which is the failure to be statistically consistent) are essentially that parallel changes exceed informative, nonparallel changes. It is possible for these methods to be inconsistent even when change is improbable a priori, provided that evolutionary rates in different lineages are sufficiently unequal. It is by extension of this approach that we may provide a sound methodology for evaluating methods of phylogenetic inference.

- ⋆ Felsenstein (1985) — confidence limits on phylogenies: the bootstrap.

  CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. Felsenstein (1985). *Evolution*. doi:10.1111/j.1558-5646.1985.tb00420.x

  > The recently-developed statistical method known as the "bootstrap" can be used to place confidence intervals on phylogenies. It involves resampling points from one's own data, with replacement, to create a series of bootstrap samples of the same size as the original data. Each of these is analyzed, and the variation among the resulting estimates taken to indicate the size of the error involved in making estimates from the original data. In the case of phylogenies, it is argued that the proper method of resampling is to keep all of the original species while sampling characters with replacement, under the assumption that the characters have been independently drawn by the systematist and have evolved independently. Majority-rule consensus trees can be used to construct a phylogeny showing all of the inferred monophyletic groups that occurred in a majority of the bootstrap samples. If a group shows up 95% of the time or more, the evidence for it is taken to be statistically significant. Existing computer programs can be used to analyze different bootstrap samples by using weights on the characters, the weight of a character being how many times it was drawn in bootstrap sampling. When all characters are perfectly compatible, as envisioned by Hennig, bootstrap sampling becomes unnecessary; the bootstrap method would show significant evidence for a group if it is defined by three or more characters.

- Studier and Keppler (1988), Gascuel (2006) — what neighbour joining does and why it works.

  A note on the neighbor-joining algorithm of Saitou and Nei. Studier & Keppler (1988). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a040527

  Neighbor-Joining Revealed. Gascuel (2006). *Molecular Biology and Evolution*. doi:10.1093/molbev/msl072

  > It is nearly 20 years since the landmark paper (Saitou and Nei 1987) in Molecular Biology and Evolution introducing Neighbor-Joining (NJ). The method has become the most widely used method for building phylogenetic trees from distances, and the original paper has been cited about 13,000 times (Science Citation Index). Yet the question "what does the NJ method seek to do?" has until recently proved somewhat elusive, leading to some imprecise claims and misunderstanding. However, a rigorous answer to this question has recently been provided by further mathematical investigation, and the purpose of this note is to highlight these results and their significance for interpreting NJ. The origins of this story lie in a paper by Pauplin (2000) though its continuation has unfolded in more mathematically inclined literature. Our aim here is to make these findings more widely accessible.

- Desper and Gascuel (2002) — FastME and the minimum-evolution criterion.

  Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. Desper & Gascuel (2002). *Journal of Computational Biology*. doi:10.1089/106652702761034136

  > The Minimum Evolution (ME) approach to phylogeny estimation has been shown to be statistically consistent when it is used in conjunction with ordinary least-squares (OLS) fitting of a metric to a tree structure. The traditional approach to using ME has been to start with the Neighbor Joining (NJ) topology for a given matrix and then do a topological search from that starting point. The first stage requires O(n(3)) time, where n is the number of taxa, while the current implementations of the second are in O(p n(3)) or more, where p is the number of swaps performed by the program. In this paper, we examine a greedy approach to minimum evolution which produces a starting topology in O(n(2)) time. Moreover, we provide an algorithm that searches for the best topology using nearest neighbor interchanges (NNIs), where the cost of doing p NNIs is O(n(2) + p n), i.e., O(n(2)) in practice because p is always much smaller

than n. The Greedy Minimum Evolution (GME) algorithm, when used in combination with NNIs, produces trees which are fairly close to NJ trees in terms of topological accuracy. We also examine ME under a balanced weighting scheme, where sibling subtrees have equal weight, as opposed to the standard "unweighted" OLS, where all taxa have the same weight so that the weight of a subtree is equal to the number of its taxa. The balanced minimum evolution scheme (BME) runs slower than the OLS version, requiring O(n(2) x diam(T)) operations to build the starting tree and O(p n x diam(T)) to perform the NNIs, where diam(T) is the topological diameter of the output tree. In the usual Yule-Harding distribution on phylogenetic trees, the diameter expectation is in log(n), so our algorithms are in practice faster that NJ. Moreover, this BME scheme yields a very significant improvement over NJ and other distance-based algorithms, especially with large trees, in terms of topological accuracy.

- Rannala and Yang (1996), Huelsenbeck et al. (2001) — Bayesian inference of phylogeny.

  Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Rannala & Yang (1996). *Journal of Molecular Evolution*. doi:10.1007/bf02338839

  > A new method is presented for inferring evolutionary trees using nucleotide sequence data. The birth-death process is used as a model of speciation and extinction to specify the prior distribution of phylogenies and branching times. Nucleotide substitution is modeled by a continuous-time Markov process. Parameters of the branching model and the substitution model are estimated by maximum likelihood. The posterior probabilities of different phylogenies are calculated and the phylogeny with the highest posterior probability is chosen as the best estimate of the evolutionary relationship among species. We refer to this as the maximum posterior probability (MAP) tree. The posterior probability provides a natural measure of the reliability of the estimated phylogeny. Two example data sets are analyzed to infer the phylogenetic relationship of human, chimpanzee, gorilla, and orangutan. The best trees estimated by the new method are the same as those from the maximum likelihood analysis of separate topologies, but the posterior probabilities are quite different from the bootstrap proportions. The results of the method are found to be insensitive to changes in the rate parameter of the branching process.

  Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Huelsenbeck et al. (2001). *Science*. doi:10.1126/science.1065889

  > As a discipline, phylogenetics is becoming transformed by a flood of molecular data. These data allow broad questions to be asked about the history of life, but also present difficult statistical and computational problems. Bayesian inference of phylogeny brings a new perspective to a number of outstanding issues in evolutionary biology, including the analysis of large phylogenetic trees and complex evolutionary models and the detection of the footprint of natural selection in DNA sequences.

B.4.3 Substitution models and support

- Jukes and Cantor (1969), Kimura (1980), Hasegawa et al. (1985) — the classical nucleotide models.

  Evolution of Protein Molecules. Jukes & Cantor (1969). *Mammalian Protein Metabolism*. doi:10.1016/b978-1-4832-3211-9.50009-7

  A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Kimura (1980). *Journal of Molecular Evolution*. doi:10.1007/bf01731581

  > Some simple formulae were obtained which enable us to estimate evolutionary distances in terms of the number of nucleotide substitutions (and, also, the evolutionary rates when the divergence times are known). In comparing a pair of nucleotide sequences, we distinguish two types of differences; if homologous sites are occupied by different nucleotide bases but both are purines or both pyrimidines, the difference is called type I (or "transition" type), while, if one of the two is a purine and the other is a pyrimidine, the difference is called type II (or "transversion" type). Letting P and Q be respectively the fractions of nucleotide sites showing type I and type II differences between two sequences compared, then the evolutionary distance per site is K = -(1/2) ln [(1-2P-Q) square root of 1-2Q]. The evolutionary rate per year is then given by k = K/(2T), where T is the time since the divergence of the two sequences. If only the third codon positions are compared, the synonymous component of the evolutionary base substitutions per site is estimated by K'S = -(1/2) ln (1-2P-Q). Also, formulae for standard errors were

obtained. Some examples were worked out using reported globin sequences to show that synonymous substitutions occur at much higher rates than amino acid-altering substitutions in evolution.

Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Hasegawa et al. (1985). *Journal of Molecular Evolution*. doi:10.1007/bf02101694

> A new statistical method for estimating divergence dates of species from DNA sequence data by a molecular clock approach is developed. This method takes into account effectively the information contained in a set of DNA sequence data. The molecular clock of mitochondrial DNA (mtDNA) was calibrated by setting the date of divergence between primates and ungulates at the Cretaceous-Tertiary boundary (65 million years ago), when the extinction of dinosaurs occurred. A generalized least-squares method was applied in fitting a model to mtDNA sequence data, and the clock gave dates of 92.3 +/- 11.7, 13.3 +/- 1.5, 10.9 +/- 1.2, 3.7 +/- 0.6, and 2.7 +/- 0.6 million years ago (where the second of each pair of numbers is the standard deviation) for the separation of mouse, gibbon, orangutan, gorilla, and chimpanzee, respectively, from the line leading to humans. Although there is some uncertainty in the clock, this dating may pose a problem for the widely believed hypothesis that the pipedal creature Australopithecus afarensis, which lived some 3.7 million years ago at Laetoli in Tanzania and at Hadar in Ethiopia, was ancestral to man and evolved after the human-ape splitting. Another likelier possibility is that mtDNA was transferred through hybridization between a proto-human and a proto-chimpanzee after the former had developed bipedalism.

- Yang (1994) — gamma-distributed rates across sites.

Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Yang (1994). *Journal of Molecular Evolution*. doi:10.1007/bf00160154

> Two approximate methods are proposed for maximum likelihood phylogenetic estimation, which allow variable rates of substitution across nucleotide sites. Three data sets with quite different characteristics were analyzed to examine empirically the performance of these methods. The first, called the "discrete gamma model," uses several categories of rates to approximate the gamma distribution, with equal probability for each category. The mean of each category is used to represent all the rates falling in the category. The performance of this method is found to be quite good, and four such categories appear to be sufficient to produce both an optimum, or near-optimum fit by the model to the data, and also an acceptable approximation to the continuous distribution. The second method, called "fixed-rates model", classifies sites into several classes according to their rates predicted assuming the star tree. Sites in different classes are then assumed to be evolving at these fixed rates when other tree topologies are evaluated. Analyses of the data sets suggest that this method can produce reasonable results, but it seems to share some properties of a least-squares pairwise comparison; for example, interior branch lengths in nonbest trees are often found to be zero. The computational requirements of the two methods are comparable to that of Felsenstein's (1981, J Mol Evol 17:368-376) model, which assumes a single rate for all the sites.

- Whelan and Goldman (2001), Le and Gascuel (2008) — empirical protein models (WAG, LG).

A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Whelan & Goldman (2001). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a003851

> Phylogenetic inference from amino acid sequence data uses mainly empirical models of amino acid replacement and is therefore dependent on those models. Two of the more widely used models, the Dayhoff and JTT models, are estimated using similar methods that can utilize large numbers of sequences from many unrelated protein families but are somewhat unsatisfactory because they rely on assumptions that may lead to systematic error and discard a large amount of the information within the sequences. The alternative method of maximum-likelihood estimation may utilize the information in the sequence data more efficiently and suffers from no systematic error, but it has previously been applicable to relatively few sequences related by a single phylogenetic tree. Here, we combine the best attributes of these two methods using an approximate maximum-likelihood method. We implemented this approach to estimate a new model of amino acid replacement from a database of globular protein sequences comprising 3,905 amino acid sequences split into 182 protein families. While the new model has an overall structure similar to those of other commonly used models, there are significant differences. The new model outperforms the Dayhoff and JTT models with respect to maximum-likelihood values for a large majority of the protein families in our database. This suggests that it provides a better overall fit to the evolutionary process in

> globular proteins and may lead to more accurate phylogenetic tree estimates. Potentially, this matrix, and the methods used to generate it, may also be useful in other areas of research, such as biological sequence database searching, sequence alignment, and protein structure prediction, for which an accurate description of amino acid replacement is required.

An Improved General Amino Acid Replacement Matrix. Le & Gascuel (2008). *Molecular Biology and Evolution*. doi:10.1093/molbev/msn067

> Amino acid replacement matrices are an essential basis of protein phylogenetics. They are used to compute substitution probabilities along phylogeny branches and thus the likelihood of the data. They are also essential in protein alignment. A number of replacement matrices and methods to estimate these matrices from protein alignments have been proposed since the seminal work of Dayhoff et al. (1972). An important advance was achieved by Whelan and Goldman (2001) and their WAG matrix, thanks to an efficient maximum likelihood estimation approach that accounts for the phylogenies of sequences within each training alignment. We further refine this method by incorporating the variability of evolutionary rates across sites in the matrix estimation and using a much larger and diverse database than BRKALN, which was used to estimate WAG. To estimate our new matrix (called LG after the authors), we use an adaptation of the XRATE software and 3,912 alignments from Pfam, comprising approximately 50,000 sequences and approximately 6.5 million residues overall. To evaluate the LG performance, we use an independent sample consisting of 59 alignments from TreeBase and randomly divide Pfam alignments into 3,412 training and 500 test alignments. The comparison with WAG and JTT shows a clear likelihood improvement. With TreeBase, we find that 1) the average Akaike information criterion gain per site is 0.25 and 0.42, when compared with WAG and JTT, respectively; 2) LG is significantly better than WAG for 38 alignments (among 59), and significantly worse with 2 alignments only; and 3) tree topologies inferred with LG, WAG, and JTT frequently differ, indicating that using LG impacts not only the likelihood value but also the output tree. Results with the test alignments from Pfam are analogous. LG and a PHYML implementation can be downloaded from http://atgc.lirmm.fr/LG.

- Lartillot and Philippe (2004) — the CAT site-heterogeneous mixture model.

A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Lartillot & Philippe (2004). *Molecular Biology and Evolution*. doi:10.1093/molbev/msh112

> Most current models of sequence evolution assume that all sites of a protein evolve under the same substitution process, characterized by a 20 x 20 substitution matrix. Here, we propose to relax this assumption by developing a Bayesian mixture model that allows the amino-acid replacement pattern at different sites of a protein alignment to be described by distinct substitution processes. Our model, named CAT, assumes the existence of distinct processes (or classes) differing by their equilibrium frequencies over the 20 residues. Through the use of a Dirichlet process prior, the total number of classes and their respective amino-acid profiles, as well as the affiliations of each site to a given class, are all free variables of the model. In this way, the CAT model is able to adapt to the complexity actually present in the data, and it yields an estimate of the substitutional heterogeneity through the posterior mean number of classes. We show that a significant level of heterogeneity is present in the substitution patterns of proteins, and that the standard one-matrix model fails to account for this heterogeneity. By evaluating the Bayes factor, we demonstrate that the standard model is outperformed by CAT on all of the data sets which we analyzed. Altogether, these results suggest that the complexity of the pattern of substitution of real sequences is better captured by the CAT model, offering the possibility of studying its impact on phylogenetic reconstruction and its connections with structure-function determinants.

- Posada and Crandall (1998), Kalyaanamoorthy et al. (2017) — model selection (ModelTest, ModelFinder).

MODELTEST: testing the model of DNA substitution.. Posada & Crandall (1998). *Bioinformatics*. doi:10.1093/bioinformatics/14.9.817

> The program MODELTEST uses log likelihood scores to establish the model of DNA evolution that best fits the data. Availability The MODELTEST package, including the source code and some documentation is available at http://bioag.byu. edu/zoology/crandall_lab/modeltest.html.

ModelFinder: fast model selection for accurate phylogenetic estimates. Kalyaanamoorthy et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4285

> Model-based molecular phylogenetics plays an important role in comparisons of genomic data, and model selection is a key step in all such analyses. We present ModelFinder, a fast model-selection method that greatly improves the accuracy of phylogenetic estimates by incorporating a model of rate heterogeneity across sites not previously considered in this context and by allowing concurrent searches of model space and tree space.

- Hoang et al. (2017), Anisimova and Gascuel (2006), Lemoine et al. (2018) — fast and alternative measures of support.

UFBoot2: Improving the Ultrafast Bootstrap Approximation. Hoang et al. (2017). *Molecular Biology and Evolution*. doi:10.1093/molbev/msx281

> The standard bootstrap (SBS), despite being computationally intensive, is widely used in maximum likelihood phylogenetic analyses. We recently proposed the ultrafast bootstrap approximation (UFBoot) to reduce computing time while achieving more unbiased branch supports than SBS under mild model violations. UFBoot has been steadily adopted as an efficient alternative to SBS and other bootstrap approaches. Here, we present UFBoot2, which substantially accelerates UFBoot and reduces the risk of overestimating branch supports due to polytomies or severe model violations. Additionally, UFBoot2 provides suitable bootstrap resampling strategies for phylogenomic data. UFBoot2 is 778 times (median) faster than SBS and 8.4 times (median) faster than RAxML rapid bootstrap on tested data sets. UFBoot2 is implemented in the IQ-TREE software package version 1.6 and freely available at http://www.iqtree.org.

Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Anisimova & Gascuel (2006). *Systematic Biology*. doi:10.1080/10635150600755453

> We revisit statistical tests for branches of evolutionary trees reconstructed upon molecular data. A new, fast, approximate likelihood-ratio test (aLRT) for branches is presented here as a competitive alternative to nonparametric bootstrap and Bayesian estimation of branch support. The aLRT is based on the idea of the conventional LRT, with the null hypothesis corresponding to the assumption that the inferred branch has length 0. We show that the LRT statistic is asymptotically distributed as a maximum of three random variables drawn from the chi(0)2 + chi(1)2 distribution. The new aLRT of interior branch uses this distribution for significance testing, but the test statistic is approximated in a slightly conservative but practical way as 2(l1- l2), i.e., double the difference between the maximum log-likelihood values corresponding to the best tree and the second best topological arrangement around the branch of interest. Such a test is fast because the log-likelihood value l2 is computed by optimizing only over the branch of interest and the four adjacent branches, whereas other parameters are fixed at their optimal values corresponding to the best ML tree. The performance of the new test was studied on simulated 4-, 12-, and 100-taxon data sets with sequences of different lengths. The aLRT is shown to be accurate, powerful, and robust to certain violations of model assumptions. The aLRT is implemented within the algorithm used by the recent fast maximum likelihood tree estimation program PHYML (Guindon and Gascuel, 2003).

Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Lemoine et al. (2018). *Nature*. doi:10.1038/s41586-018-0043-0

> Felsenstein's application of the bootstrap method to evolutionary trees is one of the most cited scientific papers of all time. The bootstrap method, which is based on resampling and replications, is used extensively to assess the robustness of phylogenetic inferences. However, increasing numbers of sequences are now available for a wide variety of species, and phylogenies based on hundreds or thousands of taxa are becoming routine. With phylogenies of this size Felsenstein's bootstrap tends to yield very low supports, especially on deep branches. Here we propose a new version of the phylogenetic bootstrap in which the presence of inferred branches in replications is measured using a gradual 'transfer' distance rather than the binary presence or absence index used in Felsenstein's original version. The resulting supports are higher and do not induce falsely supported branches. The application of our method to large mammal, HIV and simulated datasets reveals their phylogenetic signals, whereas Felsenstein's bootstrap fails to do so.

- Shimodaira (2002) — the AU test for comparing trees.

An Approximately Unbiased Test of Phylogenetic Tree Selection. Shimodaira (2002). *Systematic Biology*. doi:10.1080/10635150290069913

An approximately unbiased (AU) test that uses a newly devised multiscale bootstrap technique was developed for general hypothesis testing of regions in an attempt to reduce test bias. It was applied to maximum-likelihood tree selection for obtaining the confidence set of trees. The AU test is based on the theory of Efron et al. (Proc. Natl. Acad. Sci. USA 93:13429-13434; 1996), but the new method provides higher-order accuracy yet simpler implementation. The AU test, like the Shimodaira-Hasegawa (SH) test, adjusts the selection bias overlooked in the standard use of the bootstrap probability and Kishino-Hasegawa tests. The selection bias comes from comparing many trees at the same time and often leads to overconfidence in the wrong trees. The SH test, though safe to use, may exhibit another type of bias such that it appears conservative. Here I show that the AU test is less biased than other methods in typical cases of tree selection. These points are illustrated in a simulation study as well as in the analysis of mammalian mitochondrial protein sequences. The theoretical argument provides a simple formula that covers the bootstrap probability test, the Kishino-Hasegawa test, the AU test, and the Zharkikh-Li test. A practical suggestion is provided as to which test should be used under particular circumstances.

B.4.4 Software

- Stamatakis (2014) — RAxML.

  RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Stamatakis (2014). *Bioinformatics*. doi:10.1093/bioinformatics/btu033

  Motivation Phylogenies are increasingly used in all fields of medical and biological research. Moreover, because of the next-generation sequencing revolution, datasets used for conducting phylogenetic analyses grow at an unprecedented pace. RAxML (Randomized Axelerated Maximum Likelihood) is a popular program for phylogenetic analyses of large datasets under maximum likelihood. Since the last RAxML paper in 2006, it has been continuously maintained and extended to accommodate the increasingly growing input datasets and to serve the needs of the user community. Results I present some of the most notable new features and extensions of RAxML, such as a substantial extension of substitution models and supported data types, the introduction of SSE3, AVX and AVX2 vector intrinsics, techniques for reducing the memory requirements of the code and a plethora of operations for conducting post-analyses on sets of trees. In addition, an up-to-date 50-page user manual covering all new RAxML options is available.

- Guindon et al. (2010) — PhyML 3.0.

  New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Guindon et al. (2010). *Systematic Biology*. doi:10.1093/sysbio/syq010

  PhyML is a phylogeny software based on the maximum-likelihood principle. Early PhyML versions used a fast algorithm performing nearest neighbor interchanges to improve a reasonable starting tree topology. Since the original publication (Guindon S., Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696-704), PhyML has been widely used (>2500 citations in ISI Web of Science) because of its simplicity and a fair compromise between accuracy and speed. In the meantime, research around PhyML has continued, and this article describes the new algorithms and methods implemented in the program. First, we introduce a new algorithm to search the tree space with user-defined intensity using subtree pruning and regrafting topological moves. The parsimony criterion is used here to filter out the least promising topology modifications with respect to the likelihood function. The analysis of a large collection of real nucleotide and amino acid data sets of various sizes demonstrates the good performance of this method. Second, we describe a new test to assess the support of the data for internal branches of a phylogeny. This approach extends the recently proposed approximate likelihood-ratio test and relies on a nonparametric, Shimodaira-Hasegawa-like procedure. A detailed analysis of real alignments sheds light on the links between this new approach and the more classical nonparametric bootstrap method. Overall, our tests show that the last version (3.0) of PhyML is fast, accurate, stable, and ready to use. A Web server and binary files are available from http://www.atgc-montpellier.fr/phyml/.

- Price et al. (2010) — FastTree 2.

  FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Price et al. (2010). *PLoS ONE*. doi:10.1371/journal.pone.0009490

> Background We recently described FastTree, a tool for inferring phylogenies for alignments with up to hundreds of thousands of sequences. Here, we describe improvements to FastTree that improve its accuracy without sacrificing scalability. Methodology/principal findings Where FastTree 1 used nearest-neighbor interchanges (NNIs) and the minimum-evolution criterion to improve the tree, FastTree 2 adds minimum-evolution subtree-pruning-regrafting (SPRs) and maximum-likelihood NNIs. FastTree 2 uses heuristics to restrict the search for better trees and estimates a rate of evolution for each site (the "CAT" approximation). Nevertheless, for both simulated and genuine alignments, FastTree 2 is slightly more accurate than a standard implementation of maximum-likelihood NNIs (PhyML 3 with default settings). Although FastTree 2 is not quite as accurate as methods that use maximum-likelihood SPRs, most of the splits that disagree are poorly supported, and for large alignments, FastTree 2 is 100-1,000 times faster. FastTree 2 inferred a topology and likelihood-based local support values for 237,882 distinct 16S ribosomal RNAs on a desktop computer in 22 hours and 5.8 gigabytes of memory. Conclusions/significance FastTree 2 allows the inference of maximum-likelihood phylogenies for huge alignments. FastTree 2 is freely available at http://www.microbesonline.org/fasttree.

- Ronquist et al. (2012) — MrBayes 3.2.

  MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Ronquist et al. (2012). *Systematic Biology*. doi:10.1093/sysbio/sys029

  > Since its introduction in 2001, MrBayes has grown in popularity as a software package for Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) methods. With this note, we announce the release of version 3.2, a major upgrade to the latest official release presented in 2003. The new version provides convergence diagnostics and allows multiple analyses to be run in parallel with convergence progress monitored on the fly. The introduction of new proposals and automatic optimization of tuning parameters has improved convergence for many problems. The new version also sports significantly faster likelihood calculations through streaming single-instruction-multiple-data extensions (SSE) and support of the BEAGLE library, allowing likelihood calculations to be delegated to graphics processing units (GPUs) on compatible hardware. Speedup factors range from around 2 with SSE code to more than 50 with BEAGLE for codon problems. Checkpointing across all models allows long runs to be completed even when an analysis is prematurely terminated. New models include relaxed clocks, dating, model averaging across time-reversible substitution models, and support for hard, negative, and partial (backbone) tree constraints. Inference of species trees from gene trees is supported by full incorporation of the Bayesian estimation of species trees (BEST) algorithms. Marginal model likelihoods for Bayes factor tests can be estimated accurately across the entire model space using the stepping stone method. The new version provides more output options than previously, including samples of ancestral states, site rates, site d(N)/d(S) rations, branch rates, and node dates. A wide range of statistics on tree parameters can also be output for visualization in FigTree and compatible software.

- Suchard et al. (2018), Bouckaert et al. (2019) — BEAST and BEAST 2.

  Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Suchard et al. (2018). *Virus Evolution*. doi:10.1093/ve/vey016

  > The Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software package has become a primary tool for Bayesian phylogenetic and phylodynamic inference from genetic sequence data. BEAST unifies molecular phylogenetic reconstruction with complex discrete and continuous trait evolution, divergence-time dating, and coalescent demographic models in an efficient statistical inference engine using Markov chain Monte Carlo integration. A convenient, cross-platform, graphical user interface allows the flexible construction of complex evolutionary analyses.

  BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. Bouckaert et al. (2019). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1006650

  > Elaboration of Bayesian phylogenetic inference methods has continued at pace in recent years with major new advances in nearly all aspects of the joint modelling of evolutionary data. It is increasingly appreciated that some evolutionary questions can only be adequately answered by combining evidence from multiple independent sources of data, including genome sequences, sampling dates, phenotypic data, radiocarbon dates, fossil occurrences, and biogeographic range information among others. Including all relevant data into a single joint model is very challenging both conceptually and computationally. Advanced computational software packages that allow robust development of compatible (sub-)models which can be composed into a full model hierarchy have played a key role in these developments. Developing such software frameworks is increasingly a major scientific activity in its own right, and comes with specific

challenges, from practical software design, development and engineering challenges to statistical and conceptual modelling challenges. BEAST 2 is one such computational software platform, and was first announced over 4 years ago. Here we describe a series of major new developments in the BEAST 2 core platform and model hierarchy that have occurred since the first release of the software, culminating in the recent 2.5 release.

- Drummond et al. (2006) — relaxed clocks and dating.

  Relaxed Phylogenetics and Dating with Confidence. Drummond et al. (2006). *PLoS Biology*. doi:10.1371/journal.pbio.0040088

  > In phylogenetics, the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum. Despite their dominance in phylogenetic inference, it is evident that both are biologically unrealistic and that the real evolutionary process lies between these two extremes. Fortunately, intermediate models employing relaxed molecular clocks have been described. These models open the gate to a new field of "relaxed phylogenetics." Here we introduce a new approach to performing relaxed phylogenetic analysis. We describe how it can be used to estimate phylogenies and divergence times in the face of uncertainty in evolutionary rates and calibration times. Our approach also provides a means for measuring the clocklikeness of datasets and comparing this measure between different genes and phylogenies. We find no significant rate autocorrelation among branches in three large datasets, suggesting that autocorrelated models are not necessarily suitable for these data. In addition, we place these datasets on the continuum of clocklikeness between a strict molecular clock and the alternative unrooted extreme. Finally, we present analyses of 102 bacterial, 106 yeast, 61 plant, 99 metazoan, and 500 primate alignments. From these we conclude that our method is phylogenetically more accurate and precise than the traditional unrooted model while adding the ability to infer a timescale to evolution.

B.4.5 Alignment uncertainty and phylogenomics

- Thorne et al. (1991) — an evolutionary model for maximum-likelihood alignment (TKF91).

  An evolutionary model for maximum likelihood alignment of DNA sequences. Thorne et al. (1991). *Journal of Molecular Evolution*. doi:10.1007/bf02193625

  > Most algorithms for the alignment of biological sequences are not derived from an evolutionary model. Consequently, these alignment algorithms lack a strong statistical basis. A maximum likelihood method for the alignment of two DNA sequences is presented. This method is based upon a statistical model of DNA sequence evolution for which we have obtained explicit transition probabilities. The evolutionary model can also be used as the basis of procedures that estimate the evolutionary parameters relevant to a pair of unaligned DNA sequences. A parameter-estimation approach which takes into account all possible alignments between two sequences is introduced; the danger of estimating evolutionary parameters from a single alignment is discussed.

- Redelings and Suchard (2005) — joint Bayesian estimation of alignment and phylogeny.

  Joint Bayesian Estimation of Alignment and Phylogeny. Redelings & Suchard (2005). *Systematic Biology*. doi:10.1080/10635150590947041

  > We describe a novel model and algorithm for simultaneously estimating multiple molecular sequence alignments and the phylogenetic trees that relate the sequences. Unlike current techniques that base phylogeny estimates on a single estimate of the alignment, we take alignment uncertainty into account by considering all possible alignments. Furthermore, because the alignment and phylogeny are constructed simultaneously, a guide tree is not needed. This sidesteps the problem in which alignments created by progressive alignment are biased toward the guide tree used to generate them. Joint estimation also allows us to model rate variation between sites when estimating the alignment and to use the evidence in shared insertion/deletions (indels) to group sister taxa in the phylogeny. Our indel model makes use of affine gap penalties and considers indels of multiple letters. We make the simplifying assumption that the indel process is identical on all branches. As a result, the probability of a gap is independent of branch length. We use a Markov chain Monte Carlo (MCMC) method to sample from the posterior of the joint model, estimating the most probable alignment and tree and their support simultaneously. We describe a new MCMC transition kernel that improves our algorithm's mixing efficiency, allowing the MCMC chains to converge even when started from arbitrary alignments. Our software implementation can estimate alignment uncertainty and we describe a method for summarizing this uncertainty in a single plot.

- Wong et al. (2008) — alignment uncertainty and genomic analysis.

  Alignment Uncertainty and Genomic Analysis. Wong et al. (2008). *Science*. doi:10.1126/science.1151532

  > The statistical methods applied to the analysis of genomic data do not account for uncertainty in the sequence alignment. Indeed, the alignment is treated as an observation, and all of the subsequent inferences depend on the alignment being correct. This may not have been too problematic for many phylogenetic studies, in which the gene is carefully chosen for, among other things, ease of alignment. However, in a comparative genomics study, the same statistical methods are applied repeatedly on thousands of genes, many of which will be difficult to align. Using genomic data from seven yeast species, we show that uncertainty in the alignment can lead to several problems, including different alignment methods resulting in different conclusions.

- Talavera and Castresana (2007), Capella-Gutiérrez et al. (2009) — alignment trimming.

  Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Talavera & Castresana (2007). *Systematic Biology*. doi:10.1080/10635150701472164

  > Alignment quality may have as much impact on phylogenetic reconstruction as the phylogenetic methods used. Not only the alignment algorithm, but also the method used to deal with the most problematic alignment regions, may have a critical effect on the final tree. Although some authors remove such problematic regions, either manually or using automatic methods, in order to improve phylogenetic performance, others prefer to keep such regions to avoid losing any information. Our aim in the present work was to examine whether phylogenetic reconstruction improves after alignment cleaning or not. Using simulated protein alignments with gaps, we tested the relative performance in diverse phylogenetic analyses of the whole alignments versus the alignments with problematic regions removed with our previously developed Gblocks program. We also tested the performance of more or less stringent conditions in the selection of blocks. Alignments constructed with different alignment methods (ClustalW, Mafft, and Probcons) were used to estimate phylogenetic trees by maximum likelihood, neighbor joining, and parsimony. We show that, in most alignment conditions, and for alignments that are not too short, removal of blocks leads to better trees. That is, despite losing some information, there is an increase in the actual phylogenetic signal. Overall, the best trees are obtained by maximum-likelihood reconstruction of alignments cleaned by Gblocks. In general, a relaxed selection of blocks is better for short alignment, whereas a stringent selection is more adequate for longer ones. Finally, we show that cleaned alignments produce better topologies although, paradoxically, with lower bootstrap. This indicates that divergent and problematic alignment regions may lead, when present, to apparently better supported although, in fact, more biased topologies.

  trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Capella-Gutiérrez et al. (2009). *Bioinformatics*. doi:10.1093/bioinformatics/btp348

  > Multiple sequence alignments are central to many areas of bioinformatics. It has been shown that the removal of poorly aligned regions from an alignment increases the quality of subsequent analyses. Such an alignment trimming phase is complicated in large-scale phylogenetic analyses that deal with thousands of alignments. Here, we present trimAl, a tool for automated alignment trimming, which is especially suited for large-scale phylogenetic analyses. trimAl can consider several parameters, alone or in multiple combinations, for selecting the most reliable positions in the alignment. These include the proportion of sequences with a gap, the level of amino acid similarity and, if several alignments for the same set of sequences are provided, the level of consistency across different alignments. Moreover, trimAl can automatically select the parameters to be used in each specific alignment so that the signal-to-noise ratio is optimized. Availability trimAl has been written in C++, it is portable to all platforms. trimAl is freely available for download (http://trimal.cgenomics.org) and can be used online through the Phylemon web server (http://phylemon2.bioinfo.cipf.es/). Supplementary Material is available at http://trimal.cgenomics.org/publications.

- Rokas et al. (2003), Jeffroy et al. (2006) — the promise and pitfalls of concatenation.

  Genome-scale approaches to resolving incongruence in molecular phylogenies. Rokas et al. (2003). *Nature*. doi:10.1038/nature02053

> One of the most pervasive challenges in molecular phylogenetics is the incongruence between phylogenies obtained using different data sets, such as individual genes. To systematically investigate the degree of incongruence, and potential methods for resolving it, we screened the genome sequences of eight yeast species and selected 106 widely distributed orthologous genes for phylogenetic analyses, singly and by concatenation. Our results suggest that data sets consisting of single or a small number of concatenated genes have a significant probability of supporting conflicting topologies. By contrast, analyses of the entire data set of concatenated genes yielded a single, fully resolved species tree with maximum support. Comparable results were obtained with a concatenation of a minimum of 20 genes; substantially more genes than commonly used but a small fraction of any genome. These results have important implications for resolving branches of the tree of life.

Phylogenomics: the beginning of incongruence?. Jeffroy et al. (2006). *Trends in Genetics*. doi:10.1016/j.tig.2006.02.003

> Until recently, molecular phylogenies based on a single or few orthologous genes often yielded contradictory results. Using multiple genes in a large concatenation was proposed to end these incongruences. Here we show that single-gene phylogenies often produce incongruences, albeit ones lacking statistically significant support. By contrast, the use of different tree reconstruction methods on different partitions of the concatenated supergene leads to well-resolved, but real (i.e. statistically significant) incongruences. Gathering a large amount of data is not sufficient to produce reliable trees, given the current limitation of tree reconstruction methods, especially when the quality of data is poor. We propose that selecting only data that contain minimal nonphylogenetic signals takes full advantage of phylogenomics and markedly reduces incongruence.

- Maddison (1997), Degnan and Rosenberg (2009), Edwards (2009) — gene trees, species trees and the multispecies coalescent.

Gene Trees in Species Trees. Maddison (1997). *Systematic Biology*. doi:10.1093/sysbio/46.3.523

> Exploration of the relationship between gene trees and their containing species trees leads to consideration of how to reconstruct species trees from gene trees and of the concept of phylogeny as a cloud of gene histories. When gene copies are sampled from various species, the gene tree relating these copies might disagree with the species phylogeny. This discord can arise from horizontal transfer (including hybridization), lineage sorting, and gene duplication and extinction. Lineage sorting could also be called deep coalescence, the failure of ancestral copies to coalesce (looking backwards in time) into a common ancestral copy until deeper than previous speciation events. These events depend on various factors; for instance, deep coalescence is more likely if the branches of the species tree are short (in generations) and wide (in population size). A similar dependence on process is found in historical biogeography and host-parasite relationships. Each of the processes of discord could yield a different parsimony criterion for reconstructing the species tree from a set of gene trees: with horizontal transfer, choose the species tree that minimizes the number of transfer events; with deep coalescence, choose the tree minimizing the number of extra gene lineages that had to coexist along species lineages; with gene duplication, choose the tree minimizing duplication and/or extinction events. Maximum likelihood methods for reconstructing the species tree are also possible because coalescence theory provides the probability that a particular gene tree would occur given a species tree (with branch lengths and widths specified). In considering these issues, one is provoked to reconsider precisely what is phylogeny. Perhaps it is misleading to view some gene trees as agreeing and other gene trees as disagreeing with the species tree; rather, all of the gene trees are part of the species tree, which can be visualized like a fuzzy statistical distribution, a cloud of gene histories. Alternatively, phylogeny might be (and has been) viewed not as a history of what happened, genetically, but as a history of what could have happened, i.e., a history of changes in the probabilities of inter-breeding.

Gene tree discordance, phylogenetic inference and the multispecies coalescent. Degnan & Rosenberg (2009). *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2009.01.009

> The field of phylogenetics is entering a new era in which trees of historical relationships between species are increasingly inferred from multilocus and genomic data. A major challenge for incorporating such large amounts of data into inference of species trees is that conflicting genealogical histories often exist in different genes throughout the genome. Recent advances in genealogical modeling suggest that resolving close species relationships is not quite as simple as applying more data to the problem. Here we discuss the complexities of genealogical discordance and review the issues that new methods for multilocus species tree inference will need to address to account successfully for naturally occurring genomic variability in evolutionary histories.

IS A NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING?. Edwards (2009). *Evolution*. doi:10.1111/j.1558-5646.2008.00549.x

> The advent and maturation of algorithms for estimating species trees-phylogenetic trees that allow gene tree heterogeneity and whose tips represent lineages, populations and species, as opposed to genes-represent an exciting confluence of phylogenetics, phylogeography, and population genetics, and ushers in a new generation of concepts and challenges for the molecular systematist. In this essay I argue that to better deal with the large multilocus datasets brought on by phylogenomics, and to better align the fields of phylogeography and phylogenetics, we should embrace the primacy of species trees, not only as a new and useful practical tool for systematics, but also as a long-standing conceptual goal of systematics that, largely due to the lack of appropriate computational tools, has been eclipsed in the past few decades. I suggest that phylogenies as gene trees are a "local optimum" for systematics, and review recent advances that will bring us to the broader optimum inherent in species trees. In addition to adopting new methods of phylogenetic analysis (and ideally reserving the term "phylogeny" for species trees rather than gene trees), the new paradigm suggests shifts in a number of practices, such as sampling data to maximize not only the number of accumulated sites but also the number of independently segregating genes; routinely using coalescent or other models in computer simulations to allow gene tree heterogeneity; and understanding better the role of concatenation in influencing topologies and confidence in phylogenies. By building on the foundation laid by concepts of gene trees and coalescent theory, and by taking cues from recent trends in multilocus phylogeography, molecular systematics stands to be enriched. Many of the challenges and lessons learned for estimating gene trees will carry over to the challenge of estimating species trees, although adopting the species tree paradigm will clarify many issues (such as the nature of polytomies and the star tree paradox), raise conceptually new challenges, or provide new answers to old questions.

- Mirarab et al. (2014), Zhang et al. (2018) — ASTRAL.

ASTRAL: genome-scale coalescent-based species tree estimation. Mirarab et al. (2014). *Bioinformatics*. doi:10.1093/bioinformatics/btu462

> Motivation Species trees provide insight into basic biology, including the mechanisms of evolution and how it modifies biomolecular function and structure, biodiversity and co-evolution between genes and species. Yet, gene trees often differ from species trees, creating challenges to species tree estimation. One of the most frequent causes for conflicting topologies between gene trees and species trees is incomplete lineage sorting (ILS), which is modelled by the multi-species coalescent. While many methods have been developed to estimate species trees from multiple genes, some which have statistical guarantees under the multi-species coalescent model, existing methods are too computationally intensive for use with genome-scale analyses or have been shown to have poor accuracy under some realistic conditions. Results We present ASTRAL, a fast method for estimating species trees from multiple genes. ASTRAL is statistically consistent, can run on datasets with thousands of genes and has outstanding accuracy-improving on MP-EST and the population tree from BUCKy, two statistically consistent leading coalescent-based methods. ASTRAL is often more accurate than concatenation using maximum likelihood, except when ILS levels are low or there are too few gene trees. Availability and implementation ASTRAL is available in open source form at https://github.com/smirarab/ASTRAL/. Datasets studied in this article are available at http://www.cs.utexas.edu/users/phylo/datasets/astral. Supplementary information Supplementary data are available at Bioinformatics online.

ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. Zhang et al. (2018). *BMC Bioinformatics*. doi:10.1186/s12859-018-2129-y

> Background Evolutionary histories can be discordant across the genome, and such discordances need to be considered in reconstructing the species phylogeny. ASTRAL is one of the leading methods for inferring species trees from gene trees while accounting for gene tree discordance. ASTRAL uses dynamic programming to search for the tree that shares the maximum number of quartet topologies with input gene trees, restricting itself to a predefined set of bipartitions. Results We introduce ASTRAL-III, which substantially improves the running time of ASTRAL-II and guarantees polynomial running time as a function of both the number of species (n) and the number of genes (k). ASTRAL-III limits the bipartition constraint set (X) to grow at most linearly with n and k. Moreover, it handles polytomies more efficiently than ASTRAL-II, exploits similarities between gene trees better, and uses several techniques to avoid searching parts of the search space that are mathematically guaranteed not to include the optimal tree. The asymptotic running time of ASTRAL-III in the presence of polytomies is [Formula: see text] where D=O(nk) is the sum of degrees of all unique nodes in input trees. The running time improvements enable us to test whether contracting low support branches in gene trees improves the accuracy by

reducing noise. In extensive simulations, we show that removing branches with very low support (e.g., below 10%) improves accuracy while overly aggressive filtering is harmful. We observe on a biological avian phylogenomic dataset of 14K genes that contracting low support branches greatly improve results. Conclusions ASTRAL-III is a faster version of the ASTRAL method for phylogenetic reconstruction and can scale up to 10,000 species. With ASTRAL-III, low support branches can be removed, resulting in improved accuracy.

- Goodman et al. (1979), Szöllősi et al. (2014) — reconciliation and joint gene-tree–species-tree inference.

  Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Goodman et al. (1979). *Systematic Zoology*.

  The Inference of Gene Trees with Species Trees. Szöllősi et al. (2014). *Systematic Biology*. doi:10.1093/sysbio/syu048

  > This article reviews the various models that have been used to describe the relationships between gene trees and species trees. Molecular phylogeny has focused mainly on improving models for the reconstruction of gene trees based on sequence alignments. Yet, most phylogeneticists seek to reveal the history of species. Although the histories of genes and species are tightly linked, they are seldom identical, because genes duplicate, are lost or horizontally transferred, and because alleles can coexist in populations for periods that may span several speciation events. Building models describing the relationship between gene and species trees can thus improve the reconstruction of gene trees when a species tree is known, and vice versa. Several approaches have been proposed to solve the problem in one direction or the other, but in general neither gene trees nor species trees are known. Only a few studies have attempted to jointly infer gene trees and species trees. These models account for gene duplication and loss, transfer or incomplete lineage sorting. Some of them consider several types of events together, but none exists currently that considers the full repertoire of processes that generate gene trees along the species tree. Simulations as well as empirical studies on genomic data show that combining gene tree-species tree models with models of sequence evolution improves gene tree reconstruction. In turn, these better gene trees provide a more reliable basis for studying genome evolution or reconstructing ancestral chromosomes and ancestral gene sequences. We predict that gene tree-species tree methods that can deal with genomic data sets will be instrumental to advancing our understanding of genomic evolution.

- Huson and Bryant (2005), Bryant (2003) — phylogenetic networks and Neighbor-Net.

  Application of Phylogenetic Networks in Evolutionary Studies. Huson & Bryant (2005). *Molecular Biology and Evolution*. doi:10.1093/molbev/msj030

  > The evolutionary history of a set of taxa is usually represented by a phylogenetic tree, and this model has greatly facilitated the discussion and testing of hypotheses. However, it is well known that more complex evolutionary scenarios are poorly described by such models. Further, even when evolution proceeds in a tree-like manner, analysis of the data may not be best served by using methods that enforce a tree structure but rather by a richer visualization of the data to evaluate its properties, at least as an essential first step. Thus, phylogenetic networks should be employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved, and, even in the absence of such events, phylogenetic networks have a useful role to play. This article reviews the terminology used for phylogenetic networks and covers both split networks and reticulate networks, how they are defined, and how they can be interpreted. Additionally, the article outlines the beginnings of a comprehensive statistical framework for applying split network methods. We show how split networks can represent confidence sets of trees and introduce a conservative statistical test for whether the conflicting signal in a network is treelike. Finally, this article describes a new program, SplitsTree4, an interactive and comprehensive tool for inferring different types of phylogenetic networks from sequences, distances, and trees.

  Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Bryant (2003). *Molecular Biology and Evolution*. doi:10.1093/molbev/msh018

  > We present Neighbor-Net, a distance based method for constructing phylogenetic networks that is based on the Neighbor-Joining (NJ) algorithm of Saitou and Nei. Neighbor-Net provides a snapshot of the data that can guide more detailed analysis. Unlike split decomposition, Neighbor-Net scales well and can quickly produce detailed and informative networks for several hundred taxa. We illustrate the method by reanalyzing three published data sets: a collection of 110 highly recombinant Salmonella multi-locus

sequence typing sequences, the 135 ”African Eve” human mitochondrial sequences published by Vigilant et al., and a collection of 12 Archeal chaperonin sequences demonstrating strong evidence for gene conversion. Neighbor-Net is available as part of the SplitsTree4 software package.

B.4.6 The tree of life

- Woese and Fox (1977), Woese et al. (1990) — the archaea and the three domains.

  Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Woese & Fox (1977). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.74.11.5088

  > A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (i) the eubacteria, comprising all typical bacteria; (ii) the archaebacteria, containing methanogenic bacteria; and (iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

  Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.. Woese et al. (1990). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.87.12.4576

  > Molecular structures and sequences are generally more revealing of evolutionary relationships than are classical phenotypes (particularly so among microorganisms). Consequently, the basis for the definition of taxa has progressively shifted from the organismal to the cellular to the molecular level. Molecular comparisons show that life on this planet divides into three primary groupings, commonly known as the eubacteria, the archaebacteria, and the eukaryotes. The three are very dissimilar, the differences that separate them being of a more profound nature than the differences that separate typical kingdoms, such as animals and plants. Unfortunately, neither of the conventionally accepted views of the natural relationships among living systems--i.e., the five-kingdom taxonomy or the eukaryote-prokaryote dichotomy--reflects this primary tripartite division of the living world. To remedy this situation we propose that a formal system of organisms be established in which above the level of kingdom there exists a new taxon called a ”domain.” Life on this planet would then be seen as comprising three domains, the Bacteria, the Archaea, and the Eucarya, each containing two or more kingdoms. (The Eucarya, for example, contain Animalia, Plantae, Fungi, and a number of others yet to be defined). Although taxonomic structure within the Bacteria and Eucarya is not treated herein, Archaea is formally subdivided into the two kingdoms Euryarchaeota (encompassing the methanogens and their phenotypically diverse relatives) and Crenarchaeota (comprising the relatively tight clustering of extremely thermophilic archaebacteria, whose general phenotype appears to resemble most the ancestral phenotype of the Archaea.

- Gogarten et al. (1989), Iwabe et al. (1989) — rooting the tree of life with ancient paralogues.

  Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes.. Gogarten et al. (1989). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.86.17.6661

  > Active transport across the vacuolar components of the eukaryotic endomembrane system is energized by a specific vacuolar H+-ATPase. The amino acid sequences of the 70- and 60-kDa subunits of the vacuolar H+-ATPase are approximately equal to 25% identical to the beta and alpha subunits, respectively, of the eubacterial-type F0F1-ATPases. We now report that the same vacuolar H+-ATPase subunits are approximately equal to 50% identical to the alpha and beta subunits, respectively, of the sulfur-metabolizing Sulfolobus acidocaldarius, an archaebacterium (Archaeobacterium). Moreover, the homologue of an 88-amino acid stretch near the amino-terminal end of the 70-kDa subunit is absent from the F0F1-ATPase beta subunit but is present in the alpha subunit of Sulfolobus. Since the two types of subunits (alpha and beta subunits; 60- and 70-kDa subunits) are homologous to each other, they must have arisen by a gene duplication that occurred prior to the last common ancestor of the eubacteria, eukaryotes, and Sulfolobus. Thus, the phylogenetic tree of the subunits can be rooted at the site where the gene duplication occurred. The inferred evolutionary tree contains two main branches: a eubacterial branch and an eocyte branch that gave rise to Sulfolobus and the eukaryotic host cell. The implication is that the vacuolar H+-ATPase of eukaryotes arose by the internalization of the plasma membrane H+-ATPase of an archaebacterial-like ancestral cell.

Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes.. Iwabe et al. (1989). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.86.23.9355

> All extant organisms are though to be classified into three primary kingdoms, eubacteria, eukaryotes, and archaebacteria. The molecular evolutionary studies on the origin and evolution of archaebacteria to date have been carried out by inferring a molecular phylogenetic tree of the primary kingdoms based on comparison of a single molecule from a variety of extant species. From such comparison, it was not possible to derive the exact evolutionary relationship among the primary kingdoms, because the root of the tree could not be determined uniquely. To overcome this difficulty, we compared a pair of duplicated genes, elongation factors Tu and G, and the alpha and beta subunits of ATPase, which are thought to have diverged by gene duplication before divergence of the primary kingdoms. Using each protein pair, we inferred a composite phylogenetic tree with two clusters corresponding to different proteins, from which the evolutionary relationship of the primary kingdoms is determined uniquely. The inferred composite trees reveal that archaebacteria are more closely related to eukaryotes than to eubacteria for all the cases. By bootstrap resamplings, this relationship is reproduced with probabilities of 0.96, 0.79, 1.0, and 1.0 for elongation factors Tu and G and for ATPase subunits alpha and beta, respectively. There are also several lines of evidence for the close sequence similarity between archaebacteria and eukaryotes. Thus we propose that this tree topology represents the general evolutionary relationship among the three primary kingdoms.

- Brown et al. (2015), Parks et al. (2018) — the candidate phyla radiation and GTDB.

Unusual biology across a group comprising more than 15% of domain Bacteria. Brown et al. (2015). *Nature*. doi:10.1038/nature14486

> A prominent feature of the bacterial domain is a radiation of major lineages that are defined as candidate phyla because they lack isolated representatives. Bacteria from these phyla occur in diverse environments and are thought to mediate carbon and hydrogen cycles. Genomic analyses of a few representatives suggested that metabolic limitations have prevented their cultivation. Here we reconstructed 8 complete and 789 draft genomes from bacteria representing >35 phyla and documented features that consistently distinguish these organisms from other bacteria. We infer that this group, which may comprise >15% of the bacterial domain, has shared evolutionary history, and describe it as the candidate phyla radiation (CPR). All CPR genomes are small and most lack numerous biosynthetic pathways. Owing to divergent 16S ribosomal RNA (rRNA) gene sequences, 50-100% of organisms sampled from specific phyla would evade detection in typical cultivation-independent surveys. CPR organisms often have self-splicing introns and proteins encoded within their rRNA genes, a feature rarely reported in bacteria. Furthermore, they have unusual ribosome compositions. All are missing a ribosomal protein often absent in symbionts, and specific lineages are missing ribosomal proteins and biogenesis factors considered universal in bacteria. This implies different ribosome structures and biogenesis mechanisms, and underlines unusual biology across a large part of the bacterial domain.

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Parks et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4229

> Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

- Weiss et al. (2016), Moody et al. (2024) — the nature of LUCA.

The physiology and habitat of the last universal common ancestor. Weiss et al. (2016). *Nature Microbiology*. doi:10.1038/nmicrobiol.2016.116

> The concept of a last universal common ancestor of all cells (LUCA, or the progenote) is central to the study of early evolution and life's origin, yet information about how and where LUCA lived is lacking. We investigated all clusters and phylogenetic trees for 6.1 million protein coding genes from sequenced prokaryotic genomes in order to reconstruct the microbial ecology of LUCA. Among 286,514 protein clusters, we identified 355 protein families (~0.1%) that trace to LUCA by phylogenetic criteria. Because these proteins are not universally distributed, they can shed light on LUCA's physiology. Their functions, properties and prosthetic groups depict LUCA as anaerobic, CO2-fixing, H2-dependent with a Wood-Ljungdahl pathway, N2-fixing and thermophilic. LUCA's biochemistry was replete with FeS clusters and radical reaction mechanisms. Its cofactors reveal dependence upon transition metals, flavins, S-adenosyl methionine, coenzyme A, ferredoxin, molybdopterin, corrins and selenium. Its genetic code required nucleoside modifications and S-adenosyl methionine-dependent methylations. The 355 phylogenies identify clostridia and methanogens, whose modern lifestyles resemble that of LUCA, as basal among their respective domains. LUCA inhabited a geochemically active environment rich in H2, CO2 and iron. The data support the theory of an autotrophic origin of life involving the Wood-Ljungdahl pathway in a hydrothermal setting.

The nature of the last universal common ancestor and its impact on the early Earth system. Moody et al. (2024). *Nature Ecology & Evolution*. doi:10.1038/s41559-024-02461-1

> The nature of the last universal common ancestor (LUCA), its age and its impact on the Earth system have been the subject of vigorous debate across diverse disciplines, often based on disparate data and methods. Age estimates for LUCA are usually based on the fossil record, varying with every reinterpretation. The nature of LUCA's metabolism has proven equally contentious, with some attributing all core metabolisms to LUCA, whereas others reconstruct a simpler life form dependent on geochemistry. Here we infer that LUCA lived ~4.2 Ga (4.09-4.33 Ga) through divergence time analysis of pre-LUCA gene duplicates, calibrated using microbial fossils and isotope records under a new cross-bracing implementation. Phylogenetic reconciliation suggests that LUCA had a genome of at least 2.5 Mb (2.49-2.99 Mb), encoding around 2,600 proteins, comparable to modern prokaryotes. Our results suggest LUCA was a prokaryote-grade anaerobic acetogen that possessed an early immune system. Although LUCA is sometimes perceived as living in isolation, we infer LUCA to have been part of an established ecological system. The metabolism of LUCA would have provided a niche for other microbial community members and hydrogen recycling by atmospheric photochemistry could have supported a modestly productive early ecosystem.

- Coleman et al. (2021) — rooting the bacterial tree.

  A rooted phylogeny resolves early bacterial evolution. Coleman et al. (2021). *Science*. doi:10.1126/science.abe0511

  > A rooted bacterial tree is necessary to understand early evolution, but the position of the root is contested. Here, we model the evolution of 11,272 gene families to identify the root, extent of horizontal gene transfer (HGT), and the nature of the last bacterial common ancestor (LBCA). Our analyses root the tree between the major clades Terrabacteria and Gracilicutes and suggest that LBCA was a free-living flagellated, rod-shaped double-membraned organism. Contrary to recent proposals, our analyses reject a basal placement of the Candidate Phyla Radiation, which instead branches sister to Chloroflexota within Terrabacteria. While most gene families (92%) have evidence of HGT, overall, two-thirds of gene transmissions have been vertical, suggesting that a rooted tree provides a meaningful frame of reference for interpreting bacterial evolution.

- Spang et al. (2015), Zaremba-Niedzwiedzka et al. (2017), Imachi et al. (2020), Eme et al. (2023) — Asgard archaea and the origin of eukaryotes.

  Complex archaea that bridge the gap between prokaryotes and eukaryotes. Spang et al. (2015). *Nature*. doi:10.1038/nature14447

  > The origin of the eukaryotic cell remains one of the most contentious puzzles in modern biology. Recent studies have provided support for the emergence of the eukaryotic host cell from within the archaeal domain of life, but the identity and nature of the putative archaeal ancestor remain a subject of debate. Here we describe the discovery of 'Lokiarchaeota', a novel candidate archaeal phylum, which forms a monophyletic group with eukaryotes in phylogenomic analyses, and whose genomes encode an expanded repertoire of eukaryotic signature proteins that are suggestive of sophisticated membrane remodelling capabilities. Our results provide strong support for hypotheses in which the eukaryotic host evolved from a bona fide archaeon, and demonstrate that many components that underpin eukaryote-specific features

were already present in that ancestor. This provided the host with a rich genomic 'starter-kit' to support the increase in the cellular and genomic complexity that is characteristic of eukaryotes.

Asgard archaea illuminate the origin of eukaryotic cellular complexity. Zaremba-Niedzwiedzka et al. (2017). *Nature*. doi:10.1038/nature21031

> The origin and cellular complexity of eukaryotes represent a major enigma in biology. Current data support scenarios in which an archaeal host cell and an alphaproteobacterial (mitochondrial) endosymbiont merged together, resulting in the first eukaryotic cell. The host cell is related to Lokiarchaeota, an archaeal phylum with many eukaryotic features. The emergence of the structural complexity that characterizes eukaryotic cells remains unclear. Here we describe the 'Asgard' superphylum, a group of uncultivated archaea that, as well as Lokiarchaeota, includes Thor-, Odin- and Heimdallarchaeota. Asgard archaea affiliate with eukaryotes in phylogenomic analyses, and their genomes are enriched for proteins formerly considered specific to eukaryotes. Notably, thorarchaeal genomes encode several homologues of eukaryotic membrane-trafficking machinery components, including Sec23/24 and TRAPP domains. Furthermore, we identify thorarchaeal proteins with similar features to eukaryotic coat proteins involved in vesicle biogenesis. Our results expand the known repertoire of 'eukaryote-specific' proteins in Archaea, indicating that the archaeal host cell already contained many key components that govern eukaryotic cellular complexity.

Isolation of an archaeon at the prokaryote–eukaryote interface. Imachi et al. (2020). *Nature*. doi:10.1038/s41586-019-1916-6

> The origin of eukaryotes remains unclear 1-4 . Current data suggest that eukaryotes may have emerged from an archaeal lineage known as 'Asgard' archaea 5,6 . Despite the eukaryote-like genomic features that are found in these archaea, the evolutionary transition from archaea to eukaryotes remains unclear, owing to the lack of cultured representatives and corresponding physiological insights. Here we report the decade-long isolation of an Asgard archaeon related to Lokiarchaeota from deep marine sediment. The archaeon-'Candidatus Prometheoarchaeum syntrophicum' strain MK-D1-is an anaerobic, extremely slow-growing, small coccus (around 550 nm in diameter) that degrades amino acids through syntrophy. Although eukaryote-like intracellular complexes have been proposed for Asgard archaea 6 , the isolate has no visible organelle-like structure. Instead, Ca. P. syntrophicum is morphologically complex and has unique protrusions that are long and often branching. On the basis of the available data obtained from cultivation and genomics, and reasoned interpretations of the existing literature, we propose a hypothetical model for eukaryogenesis, termed the entangle-engulf-endogenize (also known as E 3) model.

Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Eme et al. (2023). *Nature*. doi:10.1038/s41586-023-06186-2

> In the ongoing debates about eukaryogenesis-the series of evolutionary events leading to the emergence of the eukaryotic cell from prokaryotic ancestors-members of the Asgard archaea play a key part as the closest archaeal relatives of eukaryotes 1 . However, the nature and phylogenetic identity of the last common ancestor of Asgard archaea and eukaryotes remain unresolved 2-4 . Here we analyse distinct phylogenetic marker datasets of an expanded genomic sampling of Asgard archaea and evaluate competing evolutionary scenarios using state-of-the-art phylogenomic approaches. We find that eukaryotes are placed, with high confidence, as a well-nested clade within Asgard archaea and as a sister lineage to Hodarchaeales, a newly proposed order within Heimdallarchaeia. Using sophisticated gene tree and species tree reconciliation approaches, we show that analogous to the evolution of eukaryotic genomes, genome evolution in Asgard archaea involved significantly more gene duplication and fewer gene loss events compared with other archaea. Finally, we infer that the last common ancestor of Asgard archaea was probably a thermophilic chemolithotroph and that the lineage from which eukaryotes evolved adapted to mesophilic conditions and acquired the genetic potential to support a heterotrophic lifestyle. Our work provides key insights into the prokaryote-to-eukaryote transition and a platform for better understanding the emergence of cellular complexity in eukaryotic cells.

- Williams et al. (2019) — the two-domain tree of life.

Phylogenomics provides robust support for a two-domains tree of life. Williams et al. (2019). *Nature Ecology & Evolution*. doi:10.1038/s41559-019-1040-x

> Hypotheses about the origin of eukaryotic cells are classically framed within the context of a universal 'tree of life' based on conserved core genes. Vigorous ongoing debate about eukaryote origins is based on assertions that the topology of the tree of life depends on the taxa included and the choice and quality of genomic data analysed. Here we have reanalysed the evidence underpinning those claims and apply more data to the question by using supertree and coalescent methods to interrogate >3,000 gene families in archaea and eukaryotes. We find that eukaryotes consistently originate from within the archaea in a two-domains tree when due consideration is given to the fit between model and data. Our analyses support a close relationship between eukaryotes and Asgard archaea and identify the Heimdallarchaeota as the current best candidate for the closest archaeal relatives of the eukaryotic nuclear lineage.

- Burki et al. (2020) — the new tree of eukaryotes.

  The New Tree of Eukaryotes. Burki et al. (2020). *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2019.08.008

  > For 15 years, the eukaryote Tree of Life (eToL) has been divided into five to eight major groupings, known as 'supergroups'. However, the tree has been profoundly rearranged during this time. The new eToL results from the widespread application of phylogenomics and numerous discoveries of major lineages of eukaryotes, mostly free-living heterotrophic protists. The evidence that supports the tree has transitioned from a synthesis of molecular phylogenetics and biological characters to purely molecular phylogenetics. Most current supergroups lack defining morphological or cell-biological characteristics, making the supergroup label even more arbitrary than before. Going forward, the combination of traditional culturing with maturing culture-free approaches and phylogenomics should accelerate the process of completing and resolving the eToL at its deepest levels.

- Betts et al. (2018) — dating early evolution with genomes and fossils.

  Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Betts et al. (2018). *Nature Ecology & Evolution*. doi:10.1038/s41559-018-0644-x

  > Establishing a unified timescale for the early evolution of Earth and life is challenging and mired in controversy because of the paucity of fossil evidence, the difficulty of interpreting it and dispute over the deepest branching relationships in the tree of life. Surprisingly, it remains perhaps the only episode in the history of life where literal interpretations of the fossil record hold sway, revised with every new discovery and reinterpretation. We derive a timescale of life, combining a reappraisal of the fossil material with new molecular clock analyses. We find the last universal common ancestor of cellular life to have predated the end of late heavy bombardment (>3.9 billion years ago (Ga)). The crown clades of the two primary divisions of life, Eubacteria and Archaebacteria, emerged much later (<3.4 Ga), relegating the oldest fossil evidence for life to their stem lineages. The Great Oxidation Event significantly predates the origin of modern Cyanobacteria, indicating that oxygenic photosynthesis evolved within the cyanobacterial stem lineage. Modern eukaryotes do not constitute a primary lineage of life and emerged late in Earth's history (<1.84 Ga), falsifying the hypothesis that the Great Oxidation Event facilitated their radiation. The symbiotic origin of mitochondria at 2.053-1.21 Ga reflects a late origin of the total-group Alphaproteobacteria to which the free living ancestor of mitochondria belonged.

B.4.7 Hominoid evolution

- The Chimpanzee Sequencing and Analysis Consortium (2005), Scally et al. (2012), Prado-Martinez et al. (2013) — the chimpanzee, gorilla and great-ape genomes.

  Initial sequence of the chimpanzee genome and comparison with the human genome. Sequencing & Consortium (2005). *Nature*. doi:10.1038/nature04072

  > Here we present a draft genome sequence of the common chimpanzee (Pan troglodytes). Through comparison with the human genome, we have generated a largely complete catalogue of the genetic differences that have accumulated since the human and chimpanzee species diverged from our common ancestor, constituting approximately thirty-five million single-nucleotide changes, five million insertion/deletion events, and various chromosomal rearrangements. We use this catalogue to explore the magnitude and regional variation of mutational forces shaping these two genomes, and the strength of positive and negative selection acting on their genes. In particular, we find that the patterns of evolution in human and chimpanzee protein-coding genes are highly correlated and dominated by the fixation of neutral and

slightly deleterious alleles. We also use the chimpanzee genome as an outgroup to investigate human population genetics and identify signatures of selective sweeps in recent human evolution.

Insights into hominid evolution from the gorilla genome sequence. Scally et al. (2012). *Nature*. doi:10.1038/nature10842

> Gorillas are humans' closest living relatives after chimpanzees, and are of comparable importance for the study of human origins and evolution. Here we present the assembly and analysis of a genome sequence for the western lowland gorilla, and compare the whole genomes of all extant great ape genera. We propose a synthesis of genetic and fossil evidence consistent with placing the human-chimpanzee and human-chimpanzee-gorilla speciation events at approximately 6 and 10 million years ago. In 30% of the genome, gorilla is closer to human or chimpanzee than the latter are to each other; this is rarer around coding genes, indicating pervasive selection throughout great ape evolution, and has functional consequences in gene expression. A comparison of protein coding genes reveals approximately 500 genes showing accelerated evolution on each of the gorilla, human and chimpanzee lineages, and evidence for parallel acceleration, particularly of genes involved in hearing. We also compare the western and eastern gorilla species, estimating an average sequence divergence time 1.75 million years ago, but with evidence for more recent genetic exchange and a population bottleneck in the eastern species. The use of the genome sequence in these and future analyses will promote a deeper understanding of great ape biology and evolution.

Great ape genetic diversity and population history. Prado-Martinez et al. (2013). *Nature*. doi:10.1038/nature12228

> Most great ape genetic variation remains uncharacterized; however, its study is critical for understanding population history, recombination, selection and susceptibility to disease. Here we sequence to high coverage a total of 79 wild- and captive-born individuals representing all six great ape species and seven subspecies and report 88.8 million single nucleotide polymorphisms. Our analysis provides support for genetically distinct populations within each species, signals of gene flow, and the split of common chimpanzees into two distinct groups: Nigeria-Cameroon/western and central/eastern populations. We find extensive inbreeding in almost all wild populations, with eastern gorillas being the most extreme. Inferred effective population sizes have varied radically over time in different lineages and this appears to have a profound effect on the genetic diversity at, or close to, genes in almost all species. We discover and assign 1,982 loss-of-function variants throughout the human and great ape lineages, determining that the rate of gene loss has not been different in the human branch compared to other internal branches in the great ape phylogeny. This comprehensive catalogue of great ape genome diversity provides a framework for understanding evolution and a resource for more effective management of wild and captive great ape populations.

- Yoo et al. (2025) — complete telomere-to-telomere genomes of the great apes.

  Complete sequencing of ape genomes. Yoo et al. (2025). *Nature*. doi:10.1038/s41586-025-08816-3

  > The most dynamic and repetitive regions of great ape genomes have traditionally been excluded from comparative studies 1-3 . Consequently, our understanding of the evolution of our species is incomplete. Here we present haplotype-resolved reference genomes and comparative analyses of six ape species: chimpanzee, bonobo, gorilla, Bornean orangutan, Sumatran orangutan and siamang. We achieve chromosome-level contiguity with substantial sequence accuracy (<1 error in 2.7 megabases) and completely sequence 215 gapless chromosomes telomere-to-telomere. We resolve challenging regions, such as the major histocompatibility complex and immunoglobulin loci, to provide in-depth evolutionary insights. Comparative analyses enabled investigations of the evolution and diversity of regions previously uncharacterized or incompletely studied without bias from mapping to the human reference genome. Such regions include newly minted gene families in lineage-specific segmental duplications, centromeric DNA, acrocentric chromosomes and subterminal heterochromatin. This resource serves as a comprehensive baseline for future evolutionary studies of humans and our closest living ape relatives.

- Hobolth et al. (2007) — speciation times and incomplete lineage sorting (CoalHMM).

  Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. Hobolth et al. (2007). *PLoS Genetics*. doi:10.1371/journal.pgen.0030007

> The genealogical relationship of human, chimpanzee, and gorilla varies along the genome. We develop a hidden Markov model (HMM) that incorporates this variation and relate the model parameters to population genetics quantities such as speciation times and ancestral population sizes. Our HMM is an analytically tractable approximation to the coalescent process with recombination, and in simulations we see no apparent bias in the HMM estimates. We apply the HMM to four autosomal contiguous human-chimp-gorilla-orangutan alignments comprising a total of 1.9 million base pairs. We find a very recent speciation time of human-chimp (4.1 +/- 0.4 million years), and fairly large ancestral effective population sizes (65,000 +/- 30,000 for the human-chimp ancestor and 45,000 +/- 10,000 for the human-chimp-gorilla ancestor). Furthermore, around 50% of the human genome coalesces with chimpanzee after speciation with gorilla. We also consider 250,000 base pairs of X-chromosome alignments and find an effective population size much smaller than 75% of the autosomal effective population sizes. Finally, we find that the rate of transitions between different genealogies correlates well with the region-wide present-day human recombination rate, but does not correlate with the fine-scale recombination rates and recombination hot spots, suggesting that the latter are evolutionarily transient.

- Patterson et al. (2006) — complex speciation of humans and chimpanzees.

  Genetic evidence for complex speciation of humans and chimpanzees. Patterson et al. (2006). *Nature*. doi:10.1038/nature04789

  > The genetic divergence time between two species varies substantially across the genome, conveying important information about the timing and process of speciation. Here we develop a framework for studying this variation and apply it to about 20 million base pairs of aligned sequence from humans, chimpanzees, gorillas and more distantly related primates. Human-chimpanzee genetic divergence varies from less than 84% to more than 147% of the average, a range of more than 4 million years. Our analysis also shows that human-chimpanzee speciation occurred less than 6.3 million years ago and probably more recently, conflicting with some interpretations of ancient fossils. Most strikingly, chromosome X shows an extremely young genetic divergence time, close to the genome minimum along nearly its entire length. These unexpected features would be explained if the human and chimpanzee lineages initially diverged, then later exchanged genes before separating permanently.

- Scally and Durbin (2012) — the human mutation rate and the dating of human evolution.

  Revising the human mutation rate: implications for understanding human evolution. Scally & Durbin (2012). *Nature Reviews Genetics*. doi:10.1038/nrg3295

  > It is now possible to make direct measurements of the mutation rate in modern humans using next-generation sequencing. These measurements reveal a value that is approximately half of that previously derived from fossil calibration, and this has implications for our understanding of demographic events in human evolution and other aspects of population genetics. Here, we discuss the implications of a lower-than-expected mutation rate in relation to the timescale of human evolution.

B.5 Genomics

Further papers for Chapter 5.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.5.1 Sequencing technologies

- Shendure et al. (2017) — DNA sequencing at 40: past, present and future.

  DNA sequencing at 40: past, present and future. Shendure et al. (2017). *Nature*. doi:10.1038/nature24286

> This review commemorates the 40th anniversary of DNA sequencing, a period in which we have already witnessed multiple technological revolutions and a growth in scale from a few kilobases to the first human genome, and now to millions of human and a myriad of other genomes. DNA sequencing has been extensively and creatively repurposed, including as a 'counter' for a vast range of molecular phenomena. We predict that in the long view of history, the impact of DNA sequencing will be on a par with that of the microscope.

- Maxam and Gilbert (1977), Sanger, Nicklen, et al. (1977) — the two original sequencing methods.

  A new method for sequencing DNA.. Maxam & Gilbert (1977). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.74.2.560

  > DNA can be sequenced by a chemical procedure that breaks a terminally labeled DNA molecule partially at each repetition of a base. The lengths of the labeled fragments then identify the positions of that base. We describe reactions that cleave DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. When the products of these four reactions are resolved by size, by electrophoresis on a polyacrylamide gel, the DNA sequence can be read from the pattern of radioactive bands. The technique will permit sequencing of at least 100 bases from the point of labeling.

  DNA sequencing with chain-terminating inhibitors. Sanger et al. (1977). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.74.12.5463

  > A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) J. Mol. Biol. 94, 441-448] but makes use of the 2',3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage varphiX174 and is more rapid and more accurate than either the plus or the minus method.

- Smith et al. (1986) — fluorescence detection in automated sequencing.

  Fluorescence detection in automated DNA sequence analysis. Smith et al. (1986). *Nature*. doi:10.1038/321674a0

  > We have developed a method for the partial automation of DNA sequence analysis. Fluorescence detection of the DNA fragments is accomplished by means of a fluorophore covalently attached to the oligonucleotide primer used in enzymatic DNA sequence analysis. A different coloured fluorophore is used for each of the reactions specific for the bases A, C, G and T. The reaction mixtures are combined and co-electrophoresed down a single polyacrylamide gel tube, the separated fluorescent bands of DNA are detected near the bottom of the tube, and the sequence information is acquired directly by computer.

- Ewing et al. (1998) — Phred base calling and quality scores.

  Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Ewing et al. (1998). *Genome Research*. doi:10.1101/gr.8.3.175

  > The availability of massive amounts of DNA sequence information has begun to revolutionize the practice of biology. As a result, current large-scale sequencing output, while impressive, is not adequate to keep pace with growing demand and, in particular, is far short of what will be required to obtain the 3-billion-base human genome sequence by the target date of 2005. To reach this goal, improved automation will be essential, and it is particularly important that human involvement in sequence data processing be significantly reduced or eliminated. Progress in this respect will require both improved accuracy of the data processing software and reliable accuracy measures to reduce the need for human involvement in error correction and make human review more efficient. Here, we describe one step toward that goal: a base-calling program for automated sequencer traces, phred, with improved accuracy. phred appears to be the first base-calling program to achieve a lower error rate than the ABI software, averaging 40%-50% fewer errors in the data sets examined independent of position in read, machine running conditions, or sequencing chemistry.

- Ronaghi et al. (1998), Margulies et al. (2005) — pyrosequencing and 454.

  A Sequencing Method Based on Real-Time Pyrophosphate. Ronaghi et al. (1998). *Science*. doi:10.1126/science.281.5375.363

Genome sequencing in microfabricated high-density picolitre reactors. Margulies et al. (2005). *Nature*. doi:10.1038/nature03959

> The proliferation of large-scale DNA-sequencing projects in recent years has driven a search for alternative methods to reduce time and cost. Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picolitre-scale volumes. Here we show the utility, throughput, accuracy and robustness of this system by shotgun sequencing and de novo assembly of the Mycoplasma genitalium genome with 96% coverage at 99.96% accuracy in one run of the machine.

- Bentley et al. (2008) — Illumina reversible-terminator chemistry.

Accurate whole human genome sequencing using reversible terminator chemistry. Bentley et al. (2008). *Nature*. doi:10.1038/nature07517

> DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally used long (400-800 base pair) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. Here we report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified in situ and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high-quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30x average depth of paired 35-base reads. We characterize four million single-nucleotide polymorphisms and four hundred thousand structural variants, many of which were previously unknown. Our approach is effective for accurate, rapid and economical whole-genome re-sequencing and many other biomedical applications.

- Eid et al. (2009), Wenger et al. (2019) — PacBio single-molecule sequencing and HiFi reads.

Real-Time DNA Sequencing from Single Polymerase Molecules. Eid et al. (2009). *Science*. doi:10.1126/science.1162986

> We present single-molecule, real-time sequencing data obtained from a DNA polymerase performing uninterrupted template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleoside triphosphates (dNTPs). We detected the temporal order of their enzymatic incorporation into a growing DNA strand with zero-mode waveguide nanostructure arrays, which provide optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions. Conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows continuous observation of DNA synthesis over thousands of bases without steric hindrance. The data report directly on polymerase dynamics, revealing distinct polymerization states and pause sites corresponding to DNA secondary structure. Sequence data were aligned with the known reference sequence to assay biophysical parameters of polymerization for each template position. Consensus sequences were generated from the single-molecule reads at 15-fold coverage, showing a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates.

Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. (2019). *Nature Biotechnology*. doi:10.1038/s41587-019-0217-9

> The DNA sequencing technologies in use today produce either highly accurate short reads or less-accurate long reads. We report the optimization of circular consensus sequencing (CCS) to improve the accuracy of single-molecule real-time (SMRT) sequencing (PacBio) and generate highly accurate (99.8%) long high-fidelity (HiFi) reads with an average length of 13.5 kilobases (kb). We applied our approach to sequence the well-characterized human HG002/NA24385 genome and obtained precision and recall rates of at least 99.91% for single-nucleotide variants (SNVs), 95.98% for insertions and deletions 15 megabases (Mb) and concordance of 99.997%, substantially outperforming assembly with less-accurate long reads.

- Deamer et al. (2016), M. Jain et al. (2018) — nanopore sequencing and ultra-long reads.

  Three decades of nanopore sequencing. Deamer et al. (2016). *Nature Biotechnology*. doi:10.1038/nbt.3423

  Nanopore sequencing and assembly of a human genome with ultra-long reads. Jain et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4060

  > We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. De novo assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

- Staden (1979), Lander and Waterman (1988) — the shotgun strategy and its mathematics.

  A strategy of DNA sequencing employing computer programs. Staden (1979). *Nucleic Acids Research*. doi:10.1093/nar/6.7.2601

  > With modern fast sequencing techniques and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps. This paper describes computer programs that can be used to order both sequence gel readings and clones. A method of coding for uncertainties in gel readings is described. These programs are available on request.

  Genomic mapping by fingerprinting random clones: A mathematical analysis. Lander & Waterman (1988). *Genomics*. doi:10.1016/0888-7543(88)90007-9

  > Results from physical mapping projects have recently been reported for the genomes of Escherichia coli, Saccharomyces cerevisiae, and Caenorhabditis elegans, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints. Although the basic approach is the same, there are many possible choices for the fingerprint used to characterize the clones and the rules for declaring overlap. In this paper, we derive simple formulas showing how the progress of a physical mapping project is affected by the nature of the fingerprinting scheme. Using these formulas, we discuss the analytic considerations involved in selecting an appropriate fingerprinting scheme for a particular project.

B.5.2 Genome assembly

- Nagarajan and Pop (2013) — sequence assembly demystified.

  Sequence assembly demystified. Nagarajan & Pop (2013). *Nature Reviews Genetics*. doi:10.1038/nrg3367

  > Advances in sequencing technologies and increased access to sequencing services have led to renewed interest in sequence and genome assembly. Concurrently, new applications for sequencing have emerged, including gene expression analysis, discovery of genomic variants and metagenomics, and each of these has different needs and challenges in terms of assembly. We survey the theoretical foundations that underlie modern assembly and highlight the options and practical trade-offs that need to be considered, focusing on how individual features address the needs of specific applications. We also review key software and the interplay between experimental design and efficacy of assembly.

- Anson and Myers (1999), Myers et al. (2000) — whole-genome shotgun assembly.

  Algorithms for whole genome shotgun sequencing. Anson & Myers (1999). *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB '99).*

  A Whole-Genome Assembly of Drosophila. Myers et al. (2000). *Science*. doi:10.1126/science.287.5461.2196

  > We report on the quality of a whole-genome assembly of Drosophila melanogaster and the nature of the computer algorithms that accomplished it. Three independent external data sources essentially agree with and support the assembly's sequence and ordering of contigs across the euchromatic portion of the genome. In addition, there are isolated contigs that we believe represent nonrepetitive pockets within the heterochromatin of the centromeres. Comparison with a previously sequenced 2.9- megabase region indicates that sequencing accuracy within nonrepetitive segments is greater than 99. 99% without manual curation. As such, this initial reconstruction of the Drosophila sequence should be of substantial value to the scientific community.

- ⋆ Myers (2005) — the fragment assembly string graph.

  The fragment assembly string graph. Myers (2005). *Bioinformatics*. doi:10.1093/bioinformatics/bti1114

  > We present a concept and formalism, the string graph, which represents all that is inferable about a DNA sequence from a collection of shotgun sequencing reads collected from it. We give time and space efficient algorithms for constructing a string graph given the collection of overlaps between the reads and, in particular, present a novel linear expected time algorithm for transitive reduction in this context. The result demonstrates that the decomposition of reads into kmers employed in the de Bruijn graph approach described earlier is not essential, and exposes its close connection to the unitig approach we developed at Celera. This paper is a preliminary piece giving the basic algorithm and results that demonstrate the efficiency and scalability of the method. These ideas are being used to build a next-generation whole genome assembler called BOA (Berkeley Open Assembler) that will easily scale to mammalian genomes.

- Idury and Waterman (1995), Pevzner et al. (2001) — the Eulerian path approach.

  A New Algorithm for DNA Sequence Assembly. Idury & Waterman (1995). *Journal of Computational Biology*. doi:10.1089/cmb.1995.2.291

  > Since the advent of rapid DNA sequencing methods in 1976, scientists have had the problem of inferring DNA sequences from sequenced fragments. Shotgun sequencing is a well-established biological and computational method used in practice. Many conventional algorithms for shotgun sequencing are based on the notion of pairwise fragment overlap. While shotgun sequencing infers a DNA sequence given the sequences of overlapping fragments, a recent and complementary method, called sequencing by hybridization (SBH), infers a DNA sequence given the set of oligomers that represents all subwords of some fixed length, k. In this paper, we propose a new computer algorithm for DNA sequence assembly that combines in a novel way the techniques of both shotgun and SBH methods. Based on our preliminary investigations, the algorithm promises to be very fast and practical for DNA sequence assembly.

  An Eulerian path approach to DNA fragment assembly. Pevzner et al. (2001). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.171285098

  > For the last 20 years, fragment assembly in DNA sequencing followed the ”overlap-layout-consensus” paradigm that is used in all currently available assembly tools. Although this approach proved useful in assembling clones, it faces difficulties in genomic shotgun assembly. We abandon the classical ”overlap-layout-consensus” approach in favor of a new euler algorithm that, for the first time, resolves the 20-year-old ”repeat problem” in fragment assembly. Our main result is the reduction of the fragment assembly to a variation of the classical Eulerian path problem that allows one to generate accurate solutions of large-scale sequencing problems. euler, in contrast to the celera assembler, does not mask such repeats but uses them instead as a powerful fragment assembly tool.

- Zerbino and Birney (2008), Simpson et al. (2009), Bankevich et al. (2012) — Velvet, ABySS and SPAdes.

  Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Zerbino & Birney (2008). *Genome Research*. doi:10.1101/gr.074492.107

  > We have developed a new set of algorithms, collectively called "Velvet," to manipulate de Bruijn graphs for genomic sequence assembly. A de Bruijn graph is a compact representation based on short words (k-mers) that is ideal for high coverage, very short read (25-50 bp) data sets. Applying Velvet to very short reads and paired-ends information only, one can produce contigs of significant length, up to 50-kb N50 length in simulations of prokaryotic data and 3-kb N50 on simulated mammalian BACs. When applied to real Solexa data sets without read pairs, Velvet generated contigs of approximately 8 kb in a prokaryote and 2 kb in a mammalian BAC, in close agreement with our simulated results without read-pair information. Velvet represents a new approach to assembly that can leverage very short reads in combination with read pairs to produce useful assemblies.

  ABySS: A parallel assembler for short read sequence data. Simpson et al. (2009). *Genome Research*. doi:10.1101/gr.089532.108

  > Widespread adoption of massively parallel deoxyribonucleic acid (DNA) sequencing instruments has prompted the recent development of de novo short read assembly algorithms. A common shortcoming of the available tools is their inability to efficiently assemble vast amounts of data generated from large-scale sequencing projects, such as the sequencing of individual human genomes to catalog natural genetic variation. To address this limitation, we developed ABySS (Assembly By Short Sequences), a parallelized sequence assembler. As a demonstration of the capability of our software, we assembled 3.5 billion paired-end reads from the genome of an African male publicly released by Illumina, Inc. Approximately 2.76 million contigs > or =100 base pairs (bp) in length were created with an N50 size of 1499 bp, representing 68% of the reference human genome. Analysis of these contigs identified polymorphic and novel sequences not present in the human reference assembly, which were validated by alignment to alternate human assemblies and to other primate genomes.

  SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Bankevich et al. (2012). *Journal of Computational Biology*. doi:10.1089/cmb.2012.0021

  > The lion's share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies. A major goal of single-cell genomics is to complement gene-centric metagenomic data with whole-genome assemblies of uncultivated organisms. Assembly of single-cell data is challenging because of highly non-uniform read coverage as well as elevated levels of sequencing errors and chimeric reads. We describe SPAdes, a new assembler for both single-cell and standard (multicell) assembly, and demonstrate that it improves on the recently released E+V-SC assembler (specialized for single-cell data) and on popular assemblers Velvet and SoapDeNovo (for multicell data). SPAdes generates single-cell assemblies, providing information about genomes of uncultivatable bacteria that vastly exceeds what may be obtained via traditional metagenomics studies. SPAdes is available online (http://bioinf.spbau.ru/spades). It is distributed as open source software.

- Chin et al. (2013), Koren et al. (2017), Kolmogorov et al. (2019) — long-read assembly (HGAP, Canu, Flye).

  Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Chin et al. (2013). *Nature Methods*. doi:10.1038/nmeth.2474

  > We present a hierarchical genome-assembly process (HGAP) for high-quality de novo microbial genome assemblies using only a single, long-insert shotgun DNA library in conjunction with Single Molecule, Real-Time (SMRT) DNA sequencing. Our method uses the longest reads as seeds to recruit all other reads for construction of highly accurate preassembled reads through a directed acyclic graph-based consensus procedure, which we follow with assembly using off-the-shelf long-read assemblers. In contrast to hybrid approaches, HGAP does not require highly accurate raw reads for error correction. We demonstrate efficient genome assembly for several microorganisms using as few as three SMRT Cell zero-mode waveguide arrays of sequencing and for BACs using just one SMRT Cell. Long repeat regions can be successfully resolved with this workflow. We also describe a consensus algorithm that incorporates SMRT sequencing primary quality values to produce de novo genome sequence exceeding 99.999% accuracy.

Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Koren et al. (2017). *Genome Research*. doi:10.1101/gr.215087.116

> Long-read single-molecule sequencing has revolutionized de novo genome assembly and enabled the automated reconstruction of reference-quality genomes. However, given the relatively high error rates of such technologies, efficient and accurate assembly of large repeats and closely related haplotypes remains challenging. We address these issues with Canu, a successor of Celera Assembler that is specifically designed for noisy single-molecule sequences. Canu introduces support for nanopore sequencing, halves depth-of-coverage requirements, and improves assembly continuity while simultaneously reducing runtime by an order of magnitude on large genomes versus Celera Assembler 8.2. These advances result from new overlapping and assembly algorithms, including an adaptive overlapping strategy based on tf-idf weighted MinHash and a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes. We demonstrate that Canu can reliably assemble complete microbial genomes and near-complete eukaryotic chromosomes using either Pacific Biosciences (PacBio) or Oxford Nanopore technologies and achieves a contig NG50 of >21 Mbp on both human and Drosophila melanogaster PacBio data sets. For assembly structures that cannot be linearly represented, Canu provides graph-based assembly outputs in graphical fragment assembly (GFA) format for analysis or integration with complementary phasing and scaffolding techniques. The combination of such highly resolved assembly graphs with long-range scaffolding information promises the complete and automated assembly of complex genomes.

Assembly of long, error-prone reads using repeat graphs. Kolmogorov et al. (2019). *Nature Biotechnology*. doi:10.1038/s41587-019-0072-8

> Accurate genome assembly is hampered by repetitive regions. Although long single molecule sequencing reads are better able to resolve genomic repeats than short-read data, most long-read assembly algorithms do not provide the repeat characterization necessary for producing optimal assemblies. Here, we present Flye, a long-read assembly algorithm that generates arbitrary paths in an unknown repeat graph, called disjointigs, and constructs an accurate repeat graph from these error-riddled disjointigs. We benchmark Flye against five state-of-the-art assemblers and show that it generates better or comparable assemblies, while being an order of magnitude faster. Flye nearly doubled the contiguity of the human genome assembly (as measured by the NGA50 assembly quality metric) compared with existing assemblers.

- Cheng et al. (2021) — hifiasm: haplotype-resolved assembly.

Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Cheng et al. (2021). *Nature Methods*. doi:10.1038/s41592-020-01056-5

> Haplotype-resolved de novo assembly is the ultimate solution to the study of sequence variations in a genome. However, existing algorithms either collapse heterozygous alleles into one consensus copy or fail to cleanly separate the haplotypes to produce high-quality phased assemblies. Here we describe hifiasm, a de novo assembler that takes advantage of long high-fidelity sequence reads to faithfully represent the haplotype information in a phased assembly graph. Unlike other graph-based assemblers that only aim to maintain the contiguity of one haplotype, hifiasm strives to preserve the contiguity of all haplotypes. This feature enables the development of a graph trio binning algorithm that greatly advances over standard trio binning. On three human and five nonhuman datasets, including California redwood with a ~30-Gb hexaploid genome, we show that hifiasm frequently delivers better assemblies than existing tools and consistently outperforms others on haplotype-resolved assembly.

- Burton et al. (2013) — chromosome-scale scaffolding with Hi-C.

Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Burton et al. (2013). *Nature Biotechnology*. doi:10.1038/nbt.2727

> Genomes assembled de novo from short reads are highly fragmented relative to the finished chromosomes of Homo sapiens and key model organisms generated by the Human Genome Project. To address this problem, we need scalable, cost-effective methods to obtain assemblies with chromosome-scale contiguity. Here we show that genome-wide chromatin interaction data sets, such as those generated by Hi-C, are a rich source of long-range information for assigning, ordering and orienting genomic sequences to chromosomes, including across centromeres. To exploit this finding, we developed an algorithm that uses Hi-C data for ultra-long-range scaffolding of de novo genome assemblies. We demonstrate the approach by combining shotgun fragment and short jump mate-pair sequences with Hi-C data to generate chromosome-scale de novo assemblies of the human, mouse and Drosophila genomes, achieving--for the human genome--98% accuracy in assigning scaffolds to chromosome groups and 99% accuracy in ordering

and orienting scaffolds within chromosome groups. Hi-C data can also be used to validate chromosomal translocations in cancer genomes.

- Bradnam et al. (2013) — Assemblathon 2.

  Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Bradnam et al. (2013). *Gigascience*. doi:10.1186/2047-217x-2-10

  > Background The process of generating raw genome sequence data continues to become cheaper, faster, and more accurate. However, assembly of such data into high-quality, finished genome sequences remains challenging. Many genome assembly tools are available, but they differ greatly in terms of their performance (speed, scalability, hardware requirements, acceptance of newer read technologies) and in their final output (composition of assembled sequence). More importantly, it remains largely unclear how to best assess the quality of assembled genome sequences. The Assemblathon competitions are intended to assess current state-of-the-art methods in genome assembly. Results In Assemblathon 2, we provided a variety of sequence data to be assembled for three vertebrate species (a bird, a fish, and snake). This resulted in a total of 43 submitted assemblies from 21 participating teams. We evaluated these assemblies using a combination of optical map data, Fosmid sequences, and several statistical methods. From over 100 different metrics, we chose ten key measures by which to assess the overall quality of the assemblies. Conclusions Many current genome assemblers produced useful assemblies, containing a significant representation of their genes and overall genome structure. However, the high degree of variability between the entries suggests that there is still much room for improvement in the field of genome assembly and that approaches which work well in assembling the genome of one species may not necessarily work well for another.

- Simão et al. (2015), Rhie et al. (2020) — BUSCO and Merqury.

  BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Simão et al. (2015). *Bioinformatics*. doi:10.1093/bioinformatics/btv351

  > Motivation Genomics has revolutionized biological research, but quality assessment of the resulting assembled sequences is complicated and remains mostly limited to technical measures like N50. Results We propose a measure for quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content. We implemented the assessment procedure in open-source software, with sets of Benchmarking Universal Single-Copy Orthologs, named BUSCO. Availability and implementation Software implemented in Python and datasets available for download from http://busco.ezlab.org. Contact evgeny.zdobnov@unige.ch Supplementary information Supplementary data are available at Bioinformatics online.

  Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Rhie et al. (2020). *Genome Biology*. doi:10.1186/s13059-020-02134-9

  > Recent long-read assemblies often exceed the quality and completeness of available reference genomes, making validation challenging. Here we present Merqury, a novel tool for reference-free assembly evaluation based on efficient k-mer set operations. By comparing k-mers in a de novo assembly to those found in unassembled high-accuracy reads, Merqury estimates base-level accuracy and completeness. For trios, Merqury can also evaluate haplotype-specific accuracy, completeness, phase block continuity, and switch errors. Multiple visualizations, such as k-mer spectrum plots, can be generated for evaluation. We demonstrate on both human and plant genomes that Merqury is a fast and robust method for assembly validation.

- Nurk et al. (2022) — the complete human genome (T2T).

  The complete sequence of a human genome. Nurk et al. (2022). *Science*. doi:10.1126/science.abj6987

  > Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion-base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include

all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

- Rhie et al. (2021) — reference-quality genomes for all vertebrates.

  Towards complete and error-free genome assemblies of all vertebrate species. Rhie et al. (2021). *Nature*. doi:10.1038/s41586-021-03451-0

  > High-quality and complete reference genome assemblies are fundamental for the application of genomics to biology, disease, and biodiversity conservation. However, such assemblies are available for only a few non-microbial species 1-4 . To address this issue, the international Genome 10K (G10K) consortium 5,6 has worked over a five-year period to evaluate and develop cost-effective methods for assembling highly accurate and nearly complete reference genomes. Here we present lessons learned from generating assemblies for 16 species that represent six major vertebrate lineages. We confirm that long-read sequencing technologies are essential for maximizing genome quality, and that unresolved complex repeats and haplotype heterozygosity are major sources of assembly error when not handled correctly. Our assemblies correct substantial errors, add missing sequence in some of the best historical reference genomes, and reveal biological discoveries. These include the identification of many false gene duplications, increases in gene sizes, chromosome rearrangements that are specific to lineages, a repeated independent chromosome breakpoint in bat genomes, and a canonical GC-rich pattern in protein-coding genes and their regulatory regions. Adopting these lessons, we have embarked on the Vertebrate Genomes Project (VGP), an international effort to generate high-quality, complete reference genomes for all of the roughly 70,000 extant vertebrate species and to help to enable a new era of discovery across the life sciences.

- Liao et al. (2023) — a draft human pangenome reference.

  A draft human pangenome reference. Liao et al. (2023). *Nature*. doi:10.1038/s41586-023-05896-x

  > Here the Human Pangenome Reference Consortium presents a first draft of the human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals 1 . These assemblies cover more than 99% of the expected sequence in each genome and are more than 99% accurate at the structural and base pair levels. Based on alignments of the assemblies, we generate a draft pangenome that captures known variants and haplotypes and reveals new alleles at structurally complex loci. We also add 119 million base pairs of euchromatic polymorphic sequences and 1,115 gene duplications relative to the existing reference GRCh38. Roughly 90 million of the additional base pairs are derived from structural variation. Using our draft pangenome to analyse short-read data reduced small variant discovery errors by 34% and increased the number of structural variants detected per haplotype by 104% compared with GRCh38-based workflows, which enabled the typing of the vast majority of structural variant alleles per sample.

B.5.3 Read mapping and variant calling

- Li and Durbin (2009) — BWA.

  Fast and accurate short read alignment with Burrows–Wheeler transform. Li & Durbin (2009). *Bioinformatics*. doi:10.1093/bioinformatics/btp324

  > Motivation The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals. Results We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows-Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g. from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest that BWA is approximately 10-20x faster than MAQ, while achieving similar accuracy. In addition, BWA outputs alignment in the new standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses

after the alignment can be achieved with the open source SAMtools software package. Availability http://maq.sourceforge.net.

- Li (2018) — minimap2.

  Minimap2: pairwise alignment for nucleotide sequences. Li (2018). *Bioinformatics*. doi:10.1093/bioinformatics/bty191

  > Motivation Recent advances in sequencing technologies promise ultra-long reads of ~100 kb in average, full-length mRNA or cDNA reads in high throughput and genomic contigs over 100 Mb in length. Existing alignment programs are unable or inefficient to process such data at scale, which presses for the development of new alignment algorithms. Results Minimap2 is a general-purpose alignment program to map DNA or long mRNA sequences against a large reference database. It works with accurate short reads of ≥100 bp in length, ≥1 kb genomic reads at error rate ~15%, full-length noisy Direct RNA or cDNA reads and assembly contigs or closely related full chromosomes of hundreds of megabases in length. Minimap2 does split-read alignment, employs concave gap cost for long insertions and deletions and introduces new heuristics to reduce spurious alignments. It is 3-4 times as fast as mainstream short-read mappers at comparable accuracy, and is ≥30 times faster than long-read genomic or cDNA mappers at higher accuracy, surpassing most aligners specialized in one type of alignment. Availability and implementation https://github.com/lh3/minimap2. Supplementary information Supplementary data are available at Bioinformatics online.

- Li et al. (2009), Danecek et al. (2021) — the SAM format, SAMtools and BCFtools.

  The Sequence Alignment/Map format and SAMtools. Li et al. (2009). *Bioinformatics*. doi:10.1093/bioinformatics/btp352

  > The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mbp) produced by different sequencing platforms. It is flexible in style, compact in size, efficient in random access and is the format in which alignments from the 1000 Genomes Project are released. SAMtools implements various utilities for post-processing alignments in the SAM format, such as indexing, variant caller and alignment viewer, and thus provides universal tools for processing read alignments. Availability http://samtools.sourceforge.net.

  Twelve years of SAMtools and BCFtools. Danecek et al. (2021). *GigaScience*. doi:10.1093/gigascience/giab008

  > Background SAMtools and BCFtools are widely used programs for processing and analysing high-throughput sequencing data. They include tools for file format conversion and manipulation, sorting, querying, statistics, variant calling, and effect analysis amongst other methods. Findings The first version appeared online 12 years ago and has been maintained and further developed ever since, with many new features and improvements added over the years. The SAMtools and BCFtools packages represent a unique collection of tools that have been used in numerous other software projects and countless genomic pipelines. Conclusion Both SAMtools and BCFtools are freely available on GitHub under the permissive MIT licence, free for both non-commercial and commercial use. Both packages have been installed >1 million times via Bioconda. The source code and documentation are available from https://www.htslib.org.

- Garrison et al. (2018) — graph genomes and mapping with vg.

  Variation graph toolkit improves read mapping by representing genetic variation in the reference. Garrison et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4227

  > Reference genomes guide our interpretation of DNA sequence data. However, conventional linear references represent only one version of each locus, ignoring variation in the population. Poor representation of an individual's genome sequence impacts read mapping and introduces bias. Variation graphs are bidirected DNA sequence graphs that compactly represent genetic variation across a population, including large-scale structural variation such as inversions and duplications. Previous graph genome software implementations have been limited by scalability or topological constraints. Here we present vg, a toolkit of computational methods for creating, manipulating, and using these structures as references at the scale of the human genome. vg provides an efficient approach to mapping reads onto arbitrary variation graphs using generalized compressed suffix arrays, with improved accuracy over alignment to a linear reference, and effectively removing reference bias. These capabilities make using variation graphs as references for DNA sequencing practical at a gigabase scale, or at the topological complexity of de novo assemblies.

- Li et al. (2008) — mapping quality and variant calling (MAQ).

  Mapping short DNA sequencing reads and calling variants using mapping quality scores. Li et al. (2008). *Genome Research*. doi:10.1101/gr.078212.108

  > New sequencing technologies promise a new era in the use of DNA sequence. However, some of these technologies produce very short reads, typically of a few tens of base pairs, and to use these reads effectively requires new algorithms and software. In particular, there is a major issue in efficiently aligning short reads to a reference genome and handling ambiguity or lack of accuracy in this alignment. Here we introduce the concept of mapping quality, a measure of the confidence that a read actually comes from the position it is aligned to by the mapping algorithm. We describe the software MAQ that can build assemblies by mapping shotgun short reads to a reference genome, using quality scores to derive genotype calls of the consensus sequence of a diploid genome, e.g., from a human sample. MAQ makes full use of mate-pair information and estimates the error probability of each read alignment. Error probabilities are also derived for the final genotype calls, using a Bayesian statistical model that incorporates the mapping qualities, error probabilities from the raw sequence quality scores, sampling of the two haplotypes, and an empirical model for correlated errors at a site. Both read mapping and genotype calling are evaluated on simulated data and real data. MAQ is accurate, efficient, versatile, and user-friendly. It is freely available at http://maq.sourceforge.net.

- Marth et al. (1999), Li (2011) — probabilistic SNP discovery and calling.

  A general approach to single-nucleotide polymorphism discovery. Marth et al. (1999). *Nature Genetics*.

  A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Li (2011). *Bioinformatics*. doi:10.1093/bioinformatics/btr509

  > Motivation Most existing methods for DNA sequence analysis rely on accurate sequences or genotypes. However, in applications of the next-generation sequencing (NGS), accurate genotypes may not be easily obtained (e.g. multi-sample low-coverage sequencing or somatic mutation discovery). These applications press for the development of new methods for analyzing sequence data with uncertainty. Results We present a statistical framework for calling SNPs, discovering somatic mutations, inferring population genetical parameters and performing association tests directly based on sequencing data without explicit genotyping or linkage-based imputation. On real data, we demonstrate that our method achieves comparable accuracy to alternative methods for estimating site allele count, for inferring allele frequency spectrum and for association mapping. We also highlight the necessity of using symmetric datasets for finding somatic mutations and confirm that for discovering rare events, mismapping is frequently the leading source of errors. Availability http://samtools.sourceforge.net. Contact hengli@broadinstitute.org.

- McKenna et al. (2010), DePristo et al. (2011) — GATK.

  The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. McKenna et al. (2010). *Genome Research*. doi:10.1101/gr.107524.110

  > Next-generation DNA sequencing (NGS) projects, such as the 1000 Genomes Project, are already revolutionizing our understanding of genetic variation among individuals. However, the massive data sets generated by NGS--the 1000 Genome pilot alone includes nearly five terabases--make writing feature-rich, efficient, and robust analysis tools difficult for even computationally sophisticated individuals. Indeed, many professionals are limited in the scope and the ease with which they can answer scientific questions by the complexity of accessing and manipulating the data produced by these machines. Here, we discuss our Genome Analysis Toolkit (GATK), a structured programming framework designed to ease the development of efficient and robust analysis tools for next-generation DNA sequencers using the functional programming philosophy of MapReduce. The GATK provides a small but rich set of data access patterns that encompass the majority of analysis tool needs. Separating specific analysis calculations from common data management infrastructure enables us to optimize the GATK framework for correctness, stability, and CPU and memory efficiency and to enable distributed and shared memory parallelization. We highlight the capabilities of the GATK by describing the implementation and application of robust, scale-tolerant tools like coverage calculators and single nucleotide polymorphism (SNP) calling. We conclude that the GATK programming framework enables developers and analysts to quickly and easily write efficient and robust NGS tools, many of which have already been incorporated into large-scale sequencing projects like the 1000 Genomes Project and The Cancer Genome Atlas.

A framework for variation discovery and genotyping using next-generation DNA sequencing data. DePristo et al. (2011). *Nature Genetics*. doi:10.1038/ng.806

> Recent advances in sequencing technology make it possible to comprehensively catalog genetic variation in population samples, creating a foundation for understanding human disease, ancestry and evolution. The amounts of raw data produced are prodigious, and many computational steps are required to translate this output into high-quality variant calls. We present a unified analytic framework to discover and genotype variation among multiple samples simultaneously that achieves sensitive and specific results across five sequencing technologies and three distinct, canonical experimental designs. Our process includes (i) initial read mapping; (ii) local realignment around indels; (iii) base quality score recalibration; (iv) SNP discovery and genotyping to find all potential variants; and (v) machine learning to separate true segregating variation from machine artifacts common to next-generation sequencing technologies. We here discuss the application of these tools, instantiated in the Genome Analysis Toolkit, to deep whole-genome, whole-exome capture and multi-sample low-pass (~4×) 1000 Genomes Project datasets.

- Poplin et al. (2018) — DeepVariant.

A universal SNP and small-indel variant caller using deep neural networks. Poplin et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4235

> Despite rapid advances in sequencing technologies, accurately calling genetic variants present in an individual genome from billions of short, errorful sequence reads remains challenging. Here we show that a deep convolutional neural network can call genetic variation in aligned next-generation sequencing read data by learning statistical relationships between images of read pileups around putative variant and true genotype calls. The approach, called DeepVariant, outperforms existing state-of-the-art tools. The learned model generalizes across genome builds and mammalian species, allowing nonhuman sequencing projects to benefit from the wealth of human ground-truth data. We further show that DeepVariant can learn to call variants in a variety of sequencing technologies and experimental designs, including deep whole genomes from 10X Genomics and Ion Ampliseq exomes, highlighting the benefits of using more automated and generalizable techniques for variant calling.

- Zook et al. (2019) — Genome in a Bottle benchmarks.

An open resource for accurately benchmarking small variant and reference calls. Zook et al. (2019). *Nature Biotechnology*. doi:10.1038/s41587-019-0074-6

> Benchmark small variant calls are required for developing, optimizing and assessing the performance of sequencing and bioinformatics methods. Here, as part of the Genome in a Bottle (GIAB) Consortium, we apply a reproducible, cloud-based pipeline to integrate multiple short- and linked-read sequencing datasets and provide benchmark calls for human genomes. We generate benchmark calls for one previously analyzed GIAB sample, as well as six genomes from the Personal Genome Project. These new genomes have broad, open consent, making this a 'first of its kind' resource that is available to the community for multiple downstream applications. We produce 17% more benchmark single nucleotide variations, 176% more indels and 12% larger benchmark regions than previously published GIAB benchmarks. We demonstrate that this benchmark reliably identifies errors in existing callsets and highlight challenges in interpreting performance metrics when using benchmarks that are not perfect or comprehensive. Finally, we identify strengths and weaknesses of callsets by stratifying performance according to variant type and genome context.

- The 1000 Genomes Project Consortium (2015) — the 1000 Genomes Project.

A global reference for human genetic variation. The 1000 Genomes Project Consortium (2015). *Nature*. doi:10.1038/nature15393

> The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

- Mahmoud et al. (2019), Sedlazeck et al. (2018) — structural variant calling.

  Structural variant calling: the long and the short of it. Mahmoud et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1828-7

  > Recent research into structural variants (SVs) has established their importance to medicine and molecular biology, elucidating their role in various diseases, regulation of gene expression, ethnic diversity, and large-scale chromosome evolution-giving rise to the differences within populations and among species. Nevertheless, characterizing SVs and determining the optimal approach for a given experimental design remains a computational and scientific challenge. Multiple approaches have emerged to target various SV classes, zygosities, and size ranges. Here, we review these approaches with respect to their ability to infer SVs across the full spectrum of large, complex variations and present computational methods for each approach.

  Accurate detection of complex structural variations using single-molecule sequencing. Sedlazeck et al. (2018). *Nature Methods*. doi:10.1038/s41592-018-0001-7

  > Structural variations are the greatest source of genetic variation, but they remain poorly understood because of technological limitations. Single-molecule long-read sequencing has the potential to dramatically advance the field, although high error rates are a challenge with existing methods. Addressing this need, we introduce open-source methods for long-read alignment (NGMLR; https://github.com/philres/ngmlr) and structural variant identification (Sniffles; https://github.com/fritzsedlazeck/Sniffles) that provide unprecedented sensitivity and precision for variant detection, even in repeat-rich regions and for complex nested events that can have substantial effects on human health. In several long-read datasets, including healthy and cancerous human genomes, we discovered thousands of novel variants and categorized systematic errors in short-read approaches. NGMLR and Sniffles can automatically filter false events and operate on low-coverage data, thereby reducing the high costs that have hindered the application of long reads in clinical and research settings.

B.5.4 Gene prediction

- Fickett (1982) — recognising coding regions.

  Recognition of protein coding regions in DNA sequences. Fickett (1982). *Nucleic Acids Research*. doi:10.1093/nar/10.17.5303

  > We give a test for protein coding regions which is based on simple and universal differences between protein-coding and noncoding DNA. The test is simple enough to use without a computer and is completely objective. The test has been thoroughly proven on 400,000 bases of sequence data: it misclassifies 5% of the regions tested and gives an answer of "No Opinion" one fifth of the time. We predict some new coding and noncoding regions in published sequences.

- Borodovsky and McIninch (1993) — GeneMark.

  GENMARK: Parallel gene recognition for both DNA strands. Borodovsky & McIninch (1993). *Computers & Chemistry*. doi:10.1016/0097-8485(93)85004-v

- ⋆ Lukashin and Borodovsky (1998) — GeneMark.hmm.

  GeneMark.hmm: new solutions for gene finding. Lukashin & Borodovsky (1998). *Nucleic Acids Research*. doi:10.1093/nar/26.4.1107

  > The number of completely sequenced bacterial genomes has been growing fast. There are computer methods available for finding genes but yet there is a need for more accurate algorithms. The GeneMark.hmm algorithm presented here was designed to improve the gene prediction quality in terms of finding exact gene boundaries. The idea was to embed the GeneMark models into naturally derived hidden Markov model framework with gene boundaries modeled as transitions between hidden states. We also used the specially derived ribosome binding site pattern to refine predictions of translation initiation codons. The algorithm was evaluated on several test sets including 10 complete bacterial genomes. It was shown that the new algorithm is significantly more accurate than GeneMark in exact gene prediction. Interestingly, the high gene finding accuracy was observed even in the case when Markov models of order

zero, one and two were used. We present the analysis of false positive and false negative predictions with the caution that these categories are not precisely defined if the public database annotation is used as a control.

- Delcher (1999), Hyatt et al. (2010) — Glimmer and Prodigal.

  Improved microbial gene identification with GLIMMER. Delcher (1999). *Nucleic Acids Research*. doi:10.1093/nar/27.23.4636

  > The GLIMMER system for microbial gene identification finds approximately 97-98% of all genes in a genome when compared with published annotation. This paper reports on two new results: (i) significant technical improvements to GLIMMER that improve its accuracy still further, and (ii) a comprehensive evaluation that demonstrates that the accuracy of the system is likely to be higher than previously recognized. A significant proportion of the genes missed by the system appear to be hypothetical proteins whose existence is only supported by the predictions of other programs. When the analysis is restricted to genes that have significant homology to genes in other organisms, GLIMMER misses <1% of known genes.

  Prodigal: prokaryotic gene recognition and translation initiation site identification. Hyatt et al. (2010). *BMC Bioinformatics*. doi:10.1186/1471-2105-11-119

  > Background The quality of automated gene prediction in microbial organisms has improved steadily over the past decade, but there is still room for improvement. Increasing the number of correct identifications, both of genes and of the translation initiation sites for each gene, and reducing the overall number of false positives, are all desirable goals. Results With our years of experience in manually curating genomes for the Joint Genome Institute, we developed a new gene prediction algorithm called Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm). With Prodigal, we focused specifically on the three goals of improved gene structure prediction, improved translation initiation site recognition, and reduced false positives. We compared the results of Prodigal to existing gene-finding methods to demonstrate that it met each of these objectives. Conclusion We built a fast, lightweight, open source gene prediction program called Prodigal http://compbio.ornl.gov/prodigal/. Prodigal achieved good results compared to existing methods, and we believe it will be a valuable asset to automated microbial annotation pipelines.

- Guigó et al. (1992) — GeneID: prediction of gene structure.

  Prediction of gene structure. Guigó et al. (1992). *Journal of Molecular Biology*. doi:10.1016/0022-2836(92)90130-c

  > We have developed a hierarchical rule base system for identifying genes in DNA sequences. Atomic sites (such as initiation codons, stop codons, acceptor sites and donor sites) are identified by a number of different methods and evaluated by a set of filters and rules chosen to maximize sensitivity; these are combined into higher-order gene elements (such as exons), evaluated, filtered and combined as equivalence classes into probable genes, which are evaluated and ranked. The system has been tested on an extensive collection of vertebrate genes smaller than 15,000 bases. Results obtained show that, on average, 88% of the predicted coding region for a transcription unit is actually coding, and 80% of the actual coding is correctly predicted. This will, in most applications, be sufficient for a search against protein sequence databases for the identification of probable gene function. In addition, the system provides a general test platform for both gene atomic site identification and the rules for their evaluation and assembly.

- Burge and Karlin (1997) — GENSCAN.

  Prediction of complete gene structures in human genomic DNA. Burge & Karlin (1997). *Journal of Molecular Biology*. doi:10.1006/jmbi.1997.0951

  > We introduce a general probabilistic model of the gene structure of human genomic sequences which incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. Distinct sets of model parameters are derived to account for the many substantial differences in gene density and structure observed in distinct C + G compositional regions of the human genome. In addition, new models of the donor and acceptor splice signals are described which capture potentially important dependencies between signal positions. The model is applied to the problem of gene identification in a computer program, GENSCAN, which identifies complete exon/intron structures of genes in genomic DNA. Novel features of the program include the capacity to predict multiple genes in a sequence, to deal with partial

as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. GENSCAN is shown to have substantially higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes, with 75 to 80% of exons identified exactly. The program is also capable of indicating fairly accurately the reliability of each predicted exon. Consistently high levels of accuracy are observed for sequences of differing C + G content and for distinct groups of vertebrates.

- Burset and Guigó (1996) — evaluation of gene prediction programs.

  Evaluation of Gene Structure Prediction Programs. Burset & Guigó (1996). *Genomics*. doi:10.1006/geno.1996.0298

  > We evaluate a number of computer programs designed to predict the structure of protein coding genes in genomic DNA sequences. Computational gene identification is set to play an increasingly important role in the development of the genome projects, as emphasis turns from mapping to large-scale sequencing. The evaluation presented here serves both to assess the current status of the problem and to identify the most promising approaches to ensure further progress. The programs analyzed were uniformly tested on a large set of vertebrate sequences with simple gene structure, and several measures of predictive accuracy were computed at the nucleotide, exon, and protein product levels. The results indicated that the predictive accuracy of the programs analyzed was lower than originally found. The accuracy was even lower when considering only those sequences that had recently been entered and that did not show any similarity to previously entered sequences. This indicates that the programs are overly dependent on the particularities of the examples they learn from. For most of the programs, accuracy in this test set ranged from 0.60 to 0.70 as measured by the Correlation Coefficient (where 1.0 corresponds to a perfect prediction and 0.0 is the value expected for a random prediction), and the average percentage of exons exactly identified was less than 50%. Only those programs including protein sequence database searches showed substantially greater accuracy. The accuracy of the programs was severely affected by relatively high rates of sequence errors. Since the set on which the programs were tested included only relatively short sequences with simple gene structure, the accuracy of the programs is likely to be even lower when used for large uncharacterized genomic sequences with complex structure. While in such cases, programs currently available may still be of great use in pinpointing the regions likely to contain exons, they are far from being powerful enough to elucidate its genomic structure completely.

- Stanke and Waack (2003), Korf (2004) — AUGUSTUS and SNAP.

  Gene prediction with a hidden Markov model and a new intron submodel. Stanke & Waack (2003). *Bioinformatics*. doi:10.1093/bioinformatics/btg1080

  > Motivation The problem of finding the genes in eukaryotic DNA sequences by computational methods is still not satisfactorily solved. Gene finding programs have achieved relatively high accuracy on short genomic sequences but do not perform well on longer sequences with an unknown number of genes in them. Here existing programs tend to predict many false exons. Results We have developed a new program, AUGUSTUS, for the ab initio prediction of protein coding genes in eukaryotic genomes. The program is based on a Hidden Markov Model and integrates a number of known methods and submodels. It employs a new way of modeling intron lengths. We use a new donor splice site model, a new model for a short region directly upstream of the donor splice site model that takes the reading frame into account and apply a method that allows better GC-content dependent parameter estimation. AUGUSTUS predicts on longer sequences far more human and drosophila genes accurately than the ab initio gene prediction programs we compared it with, while at the same time being more specific. Availability A web interface for AUGUSTUS and the executable program are located at http://augustus.gobics.de.

  Gene finding in novel genomes. Korf (2004). *BMC Bioinformatics*. doi:10.1186/1471-2105-5-59

  > Background Computational gene prediction continues to be an important problem, especially for genomes with little experimental data. Results I introduce the SNAP gene finder which has been designed to be easily adaptable to a variety of genomes. In novel genomes without an appropriate gene finder, I demonstrate that employing a foreign gene finder can produce highly inaccurate results, and that the most compatible parameters may not come from the nearest phylogenetic neighbor. I find that foreign gene finders are more usefully employed to bootstrap parameter estimation and that the resulting parameters can be highly accurate. Conclusion Since gene prediction is sensitive to species-specific parameters, every genome needs a dedicated gene finder.

- Holt and Yandell (2011), Brůna et al. (2021) — MAKER2 and BRAKER2.

MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. Holt & Yandell (2011). *BMC Bioinformatics*. doi:10.1186/1471-2105-12-491

> Background Second-generation sequencing technologies are precipitating major shifts with regards to what kinds of genomes are being sequenced and how they are annotated. While the first generation of genome projects focused on well-studied model organisms, many of today's projects involve exotic organisms whose genomes are largely terra incognita. This complicates their annotation, because unlike first-generation projects, there are no pre-existing 'gold-standard' gene-models with which to train gene-finders. Improvements in genome assembly and the wide availability of mRNA-seq data are also creating opportunities to update and re-annotate previously published genome annotations. Today's genome projects are thus in need of new genome annotation tools that can meet the challenges and opportunities presented by second-generation sequencing technologies. Results We present MAKER2, a genome annotation and data management tool designed for second-generation genome projects. MAKER2 is a multi-threaded, parallelized application that can process second-generation datasets of virtually any size. We show that MAKER2 can produce accurate annotations for novel genomes where training-data are limited, of low quality or even non-existent. MAKER2 also provides an easy means to use mRNA-seq data to improve annotation quality; and it can use these data to update legacy annotations, significantly improving their quality. We also show that MAKER2 can evaluate the quality of genome annotations, and identify and prioritize problematic annotations for manual review. Conclusions MAKER2 is the first annotation engine specifically designed for second-generation genome projects. MAKER2 scales to datasets of any size, requires little in the way of training data, and can use mRNA-seq data to improve annotation quality. It can also update and manage legacy genome annotation datasets.

BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. Brůna et al. (2021). *NAR Genomics and Bioinformatics*. doi:10.1093/nargab/lqaa108

> The task of eukaryotic genome annotation remains challenging. Only a few genomes could serve as standards of annotation achieved through a tremendous investment of human curation efforts. Still, the correctness of all alternative isoforms, even in the best-annotated genomes, could be a good subject for further investigation. The new BRAKER2 pipeline generates and integrates external protein support into the iterative process of training and gene prediction by GeneMark-EP+ and AUGUSTUS. BRAKER2 continues the line started by BRAKER1 where self-training GeneMark-ET and AUGUSTUS made gene predictions supported by transcriptomic data. Among the challenges addressed by the new pipeline was a generation of reliable hints to protein-coding exon boundaries from likely homologous but evolutionarily distant proteins. In comparison with other pipelines for eukaryotic genome annotation, BRAKER2 is fully automatic. It is favorably compared under equal conditions with other pipelines, e.g. MAKER2, in terms of accuracy and performance. Development of BRAKER2 should facilitate solving the task of harmonization of annotation of protein-coding genes in genomes of different eukaryotic species. However, we fully understand that several more innovations are needed in transcriptomic and proteomic technologies as well as in algorithmic development to reach the goal of highly accurate annotation of eukaryotic genomes.

- Tatusova et al. (2016) — NCBI prokaryotic genome annotation pipeline.

NCBI prokaryotic genome annotation pipeline. Tatusova et al. (2016). *Nucleic Acids Research*. doi:10.1093/nar/gkw569

> Recent technological advances have opened unprecedented opportunities for large-scale sequencing and analysis of populations of pathogenic species in disease outbreaks, as well as for large-scale diversity studies aimed at expanding our knowledge across the whole domain of prokaryotes. To meet the challenge of timely interpretation of structure, function and meaning of this vast genetic information, a comprehensive approach to automatic genome annotation is critically needed. In collaboration with Georgia Tech, NCBI has developed a new approach to genome annotation that combines alignment based methods with methods of predicting protein-coding and RNA genes and other functional elements directly from sequence. A new gene finding tool, GeneMarkS+, uses the combined evidence of protein and RNA placement by homology as an initial map of annotation to generate and modify ab initio gene predictions across the whole genome. Thus, the new NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) relies more on sequence similarity when confident comparative data are available, while it relies more on statistical predictions in the absence of external evidence. The pipeline provides a framework for generation and analysis of annotation on the full breadth of prokaryotic taxonomy. For

additional information on PGAP see https://www.ncbi.nlm.nih.gov/genome/annotation_prok/ and the NCBI Handbook, https://www.ncbi.nlm.nih.gov/books/NBK174280/.

- Stiehler et al. (2020) — Helixer: deep-learning gene prediction.

  Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Stiehler et al. (2020). *Bioinformatics*. doi:10.1093/bioinformatics/btaa1044

  > Motivation Current state-of-the-art tools for the de novo annotation of genes in eukaryotic genomes have to be specifically fitted for each species and still often produce annotations that can be improved much further. The fundamental algorithmic architecture for these tools has remained largely unchanged for about two decades, limiting learning capabilities. Here, we set out to improve the cross-species annotation of genes from DNA sequence alone with the help of deep learning. The goal is to eliminate the dependency on a closely related gene model while also improving the predictive quality in general with a fundamentally new architecture. Results We present Helixer, a framework for the development and usage of a cross-species deep learning model that improves significantly on performance and generalizability when compared to more traditional methods. We evaluate our approach by building a single vertebrate model for the base-wise annotation of 186 animal genomes and a separate land plant model for 51 plant genomes. Our predictions are shown to be much less sensitive to the length of the genome than those of a current state-of-the-art tool. We also present two novel post-processing techniques that each worked to further strengthen our annotations and show in-depth results of an RNA-Seq based comparison of our predictions. Our method does not yet produce comprehensive gene models but rather outputs base pair wise probabilities. Availability and implementation The source code of this work is available at https://github.com/weberlab-hhu/Helixer under the GNU General Public License v3.0. The trained models are available at https://doi.org/10.5281/zenodo.3974409. Supplementary information Supplementary data are available at Bioinformatics online.

B.5.5 Metagenomics

- Pace (1997), Handelsman et al. (1998) — the molecular view of microbial diversity, and the metagenome.

  A Molecular View of Microbial Diversity and the Biosphere. Pace (1997). *Science*. doi:10.1126/science.276.5313.734

  > Over three decades of molecular-phylogenetic studies, researchers have compiled an increasingly robust map of evolutionary diversification showing that the main diversity of life is microbial, distributed among three primary relatedness groups or domains: Archaea, Bacteria, and Eucarya. The general properties of representatives of the three domains indicate that the earliest life was based on inorganic nutrition and that photosynthesis and use of organic compounds for carbon and energy metabolism came comparatively later. The application of molecular-phylogenetic methods to study natural microbial ecosystems without the traditional requirement for cultivation has resulted in the discovery of many unexpected evolutionary lineages; members of some of these lineages are only distantly related to known organisms but are sufficiently abundant that they are likely to have impact on the chemistry of the biosphere.

  Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Handelsman et al. (1998). *Chemistry & Biology*. doi:10.1016/s1074-5521(98)90108-9

  > Cultured soil microorganisms have provided a rich source of natural-product chemistry. Because only a tiny fraction of soil microbes from soil are readily cultured, soil might be the greatest untapped resource for novel chemistry. The concept of cloning the metagenome to access the collective genomes and the biosynthetic machinery of soil microflora is explored here.

- Tyson et al. (2004), Venter et al. (2004) — the first shotgun metagenomes.

  Community structure and metabolism through reconstruction of microbial genomes from the environment. Tyson et al. (2004). *Nature*. doi:10.1038/nature02340

> Microbial communities are vital in the functioning of all ecosystems; however, most microorganisms are uncultivated, and their roles in natural systems are unclear. Here, using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of Leptospirillum group II and Ferroplasma type II, and partial recovery of three other genomes. This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single-nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The Leptospirillum group II genome had remarkably few nucleotide polymorphisms, despite the existence of low-abundance variants. The Ferroplasma type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment.

Environmental Genome Shotgun Sequencing of the Sargasso Sea. Venter et al. (2004). *Science*. doi:10.1126/science.1093857

> We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

- Qin et al. (2010), The Human Microbiome Project Consortium (2012), Sunagawa et al. (2015) — the gut, the human microbiome and the ocean.

A human gut microbial gene catalogue established by metagenomic sequencing. Qin et al. (2010). *Nature*. doi:10.1038/nature08821

> To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, approximately 150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

Structure, function and diversity of the healthy human microbiome. The Human Microbiome Project Consortium (2012). *Nature*. doi:10.1038/nature11234

> Studies of the human microbiome have revealed that even healthy individuals differ remarkably in the microbes that occupy habitats such as the gut, skin and vagina. Much of this diversity remains unexplained, although diet, environment, host genetics and early microbial exposure have all been implicated. Accordingly, to characterize the ecology of human-associated microbial communities, the Human Microbiome Project has analysed the largest cohort and set of distinct, clinically relevant body habitats so far. We found the diversity and abundance of each habitat's signature microbes to vary widely even among healthy subjects, with strong niche specialization both within and among individuals. The project encountered an estimated 81-99% of the genera, enzyme families and community configurations occupied by the healthy Western microbiome. Metagenomic carriage of metabolic pathways was stable among individuals despite variation in community structure, and ethnic/racial background proved to be one of the strongest associations of both pathways and microbes with clinical metadata. These results thus delineate the range of structural and functional configurations normal in the microbial communities of a healthy population, enabling future characterization of the epidemiology, ecology and translational applications of the human microbiome.

Structure and function of the global ocean microbiome. Sunagawa et al. (2015). *Science*. doi:10.1126/science.1261359

> Microbes are dominant drivers of biogeochemical processes, yet drawing a global picture of functional diversity, microbial community structure, and their ecological determinants remains a grand challenge. We analyzed 7.2 terabases of metagenomic data from 243 Tara Oceans samples from 68 locations in epipelagic and mesopelagic waters across the globe to generate an ocean microbial reference gene catalog with >40 million nonredundant, mostly novel sequences from viruses, prokaryotes, and picoeukaryotes. Using 139 prokaryote-enriched samples, containing >35,000 species, we show vertical stratification with epipelagic community composition mostly driven by temperature rather than other environmental factors or geography. We identify ocean microbial core functionality and reveal that >73% of its abundance is shared with the human gut microbiome despite the physicochemical differences between these two ecosystems.

- Callahan et al. (2016), Bolyen et al. (2019) — ASVs and QIIME 2.

DADA2: High-resolution sample inference from Illumina amplicon data. Callahan et al. (2016). *Nature Methods*. doi:10.1038/nmeth.3869

> We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected Lactobacillus crispatus variants.

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Bolyen et al. (2019). *Nature Biotechnology*. doi:10.1038/s41587-019-0209-9

- Wood and Salzberg (2014), Wood et al. (2019), Blanco-Míguez et al. (2023) — Kraken and MetaPhlAn.

Kraken: ultrafast metagenomic sequence classification using exact alignments. Wood & Salzberg (2014). *Genome Biology*. doi:10.1186/gb-2014-15-3-r46

> Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences. Previous programs designed for this task have been relatively slow and computationally expensive, forcing researchers to use faster abundance estimation programs, which only classify small subsets of metagenomic data. Using exact alignment of k-mers, Kraken achieves classification accuracy comparable to the fastest BLAST program. In its fastest mode, Kraken classifies 100 base pair reads at a rate of over 4.1 million reads per minute, 909 times faster than Megablast and 11 times faster than the abundance estimation program MetaPhlAn. Kraken is available at http://ccb.jhu.edu/software/kraken/.

Improved metagenomic analysis with Kraken 2. Wood et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1891-0

> Although Kraken's k-mer-based approach provides a fast taxonomic classification of metagenomic sequence data, its large memory requirements can be limiting for some applications. Kraken 2 improves upon Kraken 1 by reducing memory usage by 85%, allowing greater amounts of reference genomic data to be used, while maintaining high accuracy and increasing speed fivefold. Kraken 2 also introduces a translated search mode, providing increased sensitivity in viral metagenomics analysis.

Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Blanco-Míguez et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-023-01688-w

> Metagenomic assembly enables new organism discovery from microbial communities, but it can only capture few abundant organisms from most metagenomes. Here we present MetaPhlAn 4, which integrates information from metagenome assemblies and microbial isolate genomes for more comprehensive metagenomic taxonomic profiling. From a curated collection of 1.01 M prokaryotic reference and metagenome-assembled genomes, we define unique marker genes for 26,970 species-level genome bins, 4,992 of them taxonomically unidentified at the species level. MetaPhlAn 4 explains ~20% more reads

> in most international human gut microbiomes and >40% in less-characterized environments such as the rumen microbiome and proves more accurate than available alternatives on synthetic evaluations while also reliably quantifying organisms with no cultured isolates. Application of the method to >24,500 metagenomes highlights previously undetected species to be strong biomarkers for host conditions and lifestyles in human and mouse microbiomes and shows that even previously uncharacterized species can be genetically profiled at the resolution of single microbial strains.

- Nurk et al. (2017), Li et al. (2015) — metagenome assembly.

  metaSPAdes: a new versatile metagenomic assembler. Nurk et al. (2017). *Genome Research*. doi:10.1101/gr.213959.116

  > While metagenomics has emerged as a technology of choice for analyzing bacterial populations, the assembly of metagenomic data remains challenging, thus stifling biological discoveries. Moreover, recent studies revealed that complex bacterial populations may be composed from dozens of related strains, thus further amplifying the challenge of metagenomic assembly. metaSPAdes addresses various challenges of metagenomic assembly by capitalizing on computational ideas that proved to be useful in assemblies of single cells and highly polymorphic diploid genomes. We benchmark metaSPAdes against other state-of-the-art metagenome assemblers and demonstrate that it results in high-quality assemblies across diverse data sets.

  MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Li et al. (2015). *Bioinformatics*. doi:10.1093/bioinformatics/btv033

  > MEGAHIT is a NGS de novo assembler for assembling large and complex metagenomics data in a time- and cost-efficient manner. It finished assembling a soil metagenomics dataset with 252 Gbps in 44.1 and 99.6 h on a single computing node with and without a graphics processing unit, respectively. MEGAHIT assembles the data as a whole, i.e. no pre-processing like partitioning and normalization was needed. When compared with previous methods on assembling the soil data, MEGAHIT generated a three-time larger assembly, with longer contig N50 and average contig length; furthermore, 55.8% of the reads were aligned to the assembly, giving a fourfold improvement.

- Kang et al. (2019), Parks et al. (2015), Bowers et al. (2017) — binning, MAG quality and standards.

  MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. Kang et al. (2019). *PeerJ*. doi:10.7717/peerj.7359

  > We previously reported on MetaBAT, an automated metagenome binning software tool to reconstruct single genomes from microbial communities for subsequent analyses of uncultivated microbial species. MetaBAT has become one of the most popular binning tools largely due to its computational efficiency and ease of use, especially in binning experiments with a large number of samples and a large assembly. MetaBAT requires users to choose parameters to fine-tune its sensitivity and specificity. If those parameters are not chosen properly, binning accuracy can suffer, especially on assemblies of poor quality. Here, we developed MetaBAT 2 to overcome this problem. MetaBAT 2 uses a new adaptive binning algorithm to eliminate manual parameter tuning. We also performed extensive software engineering optimization to increase both computational and memory efficiency. Comparing MetaBAT 2 to alternative software tools on over 100 real world metagenome assemblies shows superior accuracy and computing speed. Binning a typical metagenome assembly takes only a few minutes on a single commodity workstation. We therefore recommend the community adopts MetaBAT 2 for their metagenome binning experiments. MetaBAT 2 is open source software and available at https://bitbucket.org/berkeleylab/metabat.

  CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Parks et al. (2015). *Genome Research*. doi:10.1101/gr.186072.114

  > Large-scale recovery of genomes from isolates, single cells, and metagenomic data has been made possible by advances in computational methods and substantial reductions in sequencing costs. Although this increasing breadth of draft genomes is providing key information regarding the evolutionary and functional diversity of microbial life, it has become impractical to finish all available reference genomes. Making robust biological inferences from draft genomes requires accurate estimates of their completeness and contamination. Current methods for assessing genome quality are ad hoc and generally make use of a limited number of "marker" genes conserved across all bacterial or archaeal genomes. Here we introduce CheckM, an automated method for assessing the quality of a genome using a broader set of marker genes

specific to the position of a genome within a reference genome tree and information about the collocation of these genes. We demonstrate the effectiveness of CheckM using synthetic data and a wide range of isolate-, single-cell-, and metagenome-derived genomes. CheckM is shown to provide accurate estimates of genome completeness and contamination and to outperform existing approaches. Using CheckM, we identify a diverse range of errors currently impacting publicly available isolate genomes and demonstrate that genomes obtained from single cells and metagenomic data vary substantially in quality. In order to facilitate the use of draft genomes, we propose an objective measure of genome quality that can be used to select genomes suitable for specific gene- and genome-centric analyses of microbial communities.

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Bowers et al. (2017). *Nature Biotechnology*. doi:10.1038/nbt.3893

We present two standards developed by the Genomic Standards Consortium (GSC) for reporting bacterial and archaeal genome sequences. Both are extensions of the Minimum Information about Any (x) Sequence (MIxS). The standards are the Minimum Information about a Single Amplified Genome (MISAG) and the Minimum Information about a Metagenome-Assembled Genome (MIMAG), including, but not limited to, assembly quality, and estimates of genome completeness and contamination. These standards can be used in combination with other GSC checklists, including the Minimum Information about a Genome Sequence (MIGS), Minimum Information about a Metagenomic Sequence (MIMS), and Minimum Information about a Marker Gene Sequence (MIMARKS). Community-wide adoption of MISAG and MIMAG will facilitate more robust comparative genomic analyses of bacterial and archaeal diversity.

- Almeida et al. (2019), Pasolli et al. (2019), Nayfach et al. (2020) — large catalogues of MAGs.

A new genomic blueprint of the human gut microbiota. Almeida et al. (2019). *Nature*. doi:10.1038/s41586-019-0965-1

The composition of the human gut microbiota is linked to health and disease, but knowledge of individual microbial species is needed to decipher their biological roles. Despite extensive culturing and sequencing efforts, the complete bacterial repertoire of the human gut microbiota remains undefined. Here we identify 1,952 uncultured candidate bacterial species by reconstructing 92,143 metagenome-assembled genomes from 11,850 human gut microbiomes. These uncultured genomes substantially expand the known species repertoire of the collective human gut microbiota, with a 281% increase in phylogenetic diversity. Although the newly identified species are less prevalent in well-studied populations compared to reference isolate genomes, they improve classification of understudied African and South American samples by more than 200%. These candidate species encode hundreds of newly identified biosynthetic gene clusters and possess a distinctive functional capacity that might explain their elusive nature. Our work expands the known diversity of uncultured gut bacteria, which provides unprecedented resolution for taxonomic and functional characterization of the intestinal microbiota.

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Pasolli et al. (2019). *Cell*. doi:10.1016/j.cell.2019.01.001

The body-wide human microbiome plays a role in health, but its full diversity remains uncharacterized, particularly outside of the gut and in international populations. We leveraged 9,428 metagenomes to reconstruct 154,723 microbial genomes (45% of high quality) spanning body sites, ages, countries, and lifestyles. We recapitulated 4,930 species-level genome bins (SGBs), 77% without genomes in public repositories (unknown SGBs [uSGBs]). uSGBs are prevalent (in 93% of well-assembled samples), expand underrepresented phyla, and are enriched in non-Westernized populations (40% of the total SGBs). We annotated 2.85 M genes in SGBs, many associated with conditions including infant development (94,000) or Westernization (106,000). SGBs and uSGBs permit deeper microbiome analyses and increase the average mappability of metagenomic reads from 67.76% to 87.51% in the gut (median 94.26%) and 65.14% to 82.34% in the mouth. We thus identify thousands of microbial genomes from yet-to-be-named species, expand the pangenomes of human-associated microbes, and allow better exploitation of metagenomic technologies.

A genomic catalog of Earth's microbiomes. Nayfach et al. (2020). *Nature Biotechnology*. doi:10.1038/s41587-020-0718-6

> The reconstruction of bacterial and archaeal genomes from shotgun metagenomes has enabled insights into the ecology and evolution of environmental and host-associated microbiomes. Here we applied this approach to >10,000 metagenomes collected from diverse habitats covering all of Earth's continents and oceans, including metagenomes from human and animal hosts, engineered environments, and natural and agricultural soils, to capture extant microbial, metabolic and functional potential. This comprehensive catalog includes 52,515 metagenome-assembled genomes representing 12,556 novel candidate species-level operational taxonomic units spanning 135 phyla. The catalog expands the known phylogenetic diversity of bacteria and archaea by 44% and is broadly available for streamlined comparative analyses, interactive exploration, metabolic modeling and bulk download. We demonstrate the utility of this collection for understanding secondary-metabolite biosynthetic potential and for resolving thousands of new host linkages to uncultivated viruses. This resource underscores the value of genome-centric approaches for revealing genomic properties of uncultivated microorganisms that affect ecosystem processes.

- Sczyrba et al. (2017), Meyer et al. (2022) — CAMI benchmarks.

  Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Sczyrba et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4458

  > Methods for assembly, taxonomic profiling and binning are key to interpreting metagenome data, but a lack of consensus about benchmarking complicates performance assessment. The Critical Assessment of Metagenome Interpretation (CAMI) challenge has engaged the global developer community to benchmark their programs on highly complex and realistic data sets, generated from ~700 newly sequenced microorganisms and ~600 novel viruses and plasmids and representing common experimental setups. Assembly and genome binning programs performed well for species represented by individual genomes but were substantially affected by the presence of related strains. Taxonomic profiling and binning programs were proficient at high taxonomic ranks, with a notable performance decrease below family level. Parameter settings markedly affected performance, underscoring their importance for program reproducibility. The CAMI results highlight current challenges but also provide a roadmap for software selection to answer specific research questions.

  Critical Assessment of Metagenome Interpretation: the second round of challenges. Meyer et al. (2022). *Nature Methods*. doi:10.1038/s41592-022-01431-4

  > Evaluating metagenomic software is key for optimizing metagenome interpretation and focus of the Initiative for the Critical Assessment of Metagenome Interpretation (CAMI). The CAMI II challenge engaged the community to assess methods on realistic and complex datasets with long- and short-read sequences, created computationally from around 1,700 new and known genomes, as well as 600 new plasmids and viruses. Here we analyze 5,002 results by 76 program versions. Substantial improvements were seen in assembly, some due to long-read data. Related strains still were challenging for assembly and genome recovery through binning, as was assembly quality for the latter. Profilers markedly matured, with taxon profilers and binners excelling at higher bacterial ranks, but underperforming for viruses and Archaea. Clinical pathogen detection results revealed a need to improve reproducibility. Runtime and memory usage analyses identified efficient programs, including top performers with other metrics. The results identify challenges and guide researchers in selecting methods for analyses.

B.6 Genomes

Further papers for Chapter 6.

In the HTML version, click *Abstract* to show the abstract.

B.6.1 Genome projects

- Fleischmann et al. (1995), Fraser et al. (1995) — the first bacterial genomes.

  Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. Fleischmann et al. (1995). *Science*. doi:10.1126/science.7542800

> An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium Haemophilus influenzae Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The H. influenzae Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

The Minimal Gene Complement of Mycoplasma genitalium. Fraser et al. (1995). *Science*. doi:10.1126/science.270.5235.397

> The complete nucleotide sequence (580,070 base pairs) of the Mycoplasma genitalium genome, the smallest known genome of any free-living organism, has been determined by whole-genome random sequencing and assembly. A total of only 470 predicted coding regions were identified that include genes required for DNA replication, transcription and translation, DNA repair, cellular transport, and energy metabolism. Comparison of this genome to that of Haemophilus influenzae suggests that differences in genome content are reflected as profound differences in physiology and metabolic capacity between these two organisms.

- Goffeau et al. (1996) — yeast: life with 6,000 genes.

Life with 6000 Genes. Goffeau et al. (1996). *Science*. doi:10.1126/science.274.5287.546

> The genome of the yeast Saccharomyces cerevisiae has been completely sequenced through a worldwide collaboration. The sequence of 12,068 kilobases defines 5885 potential protein-encoding genes, approximately 140 genes specifying ribosomal RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. In addition, the complete sequence provides information about the higher order organization of yeast's 16 chromosomes and allows some insight into their evolutionary history. The genome shows a considerable amount of apparent genetic redundancy, and one of the major problems to be tackled during the next stage of the yeast genome project is to elucidate the biological functions of all of these genes.

- Blattner et al. (1997) — *Escherichia coli*.

The Complete Genome Sequence of Escherichia coli K-12. Blattner et al. (1997). *Science*. doi:10.1126/science.277.5331.1453

> The 4,639,221-base pair sequence of Escherichia coli K-12 is presented. Of 4288 protein-coding genes annotated, 38 percent have no attributed function. Comparison with five other sequenced microbes reveals ubiquitous as well as narrowly distributed gene families; many families of similar genes within E. coli are also evident. The largest family of paralogous proteins contains 80 ABC transporters. The genome as a whole is strikingly organized with respect to the local direction of replication; guanines, oligonucleotides possibly related to replication and recombination, and most genes are so oriented. The genome also contains insertion sequence (IS) elements, phage remnants, and many other patches of unusual composition indicating genome plasticity through horizontal transfer.

- The C. elegans Sequencing Consortium (1998), Adams et al. (2000), The Arabidopsis Genome Initiative (2000) — the first animal and plant genomes.

Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. The C. elegans Sequencing Consortium (1998). *Science*. doi:10.1126/science.282.5396.2012

> The 97-megabase genomic sequence of the nematode Caenorhabditis elegans reveals over 19,000 genes. More than 40 percent of the predicted protein products find significant matches in other organisms. There is a variety of repeated sequences, both local and dispersed. The distinctive distribution of some repeats and highly conserved genes provides evidence for a regional organization of the chromosomes.

The Genome Sequence of Drosophila melanogaster. Adams et al. (2000). *Science*. doi:10.1126/science.287.5461.2185

> The fly Drosophila melanogaster is one of the most intensively studied organisms in biology and serves as a model system for the investigation of many developmental and cellular processes common to higher eukaryotes, including humans. We have determined the nucleotide sequence of nearly all of the approximately 120-megabase euchromatic portion of the Drosophila genome using a whole-genome shotgun sequencing strategy supported by extensive clone-based sequence and a high-quality bacterial artificial chromosome physical map. Efforts are under way to close the remaining gaps; however, the

sequence is of sufficient accuracy and contiguity to be declared substantially complete and to support an initial analysis of genome structure and preliminary gene annotation and interpretation. The genome encodes approximately 13,600 genes, somewhat fewer than the smaller Caenorhabditis elegans genome, but with comparable functional diversity.

Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. The Arabidopsis Genome Initiative (2000). *Nature*. doi:10.1038/35048692

> The flowering plant Arabidopsis thaliana is an important model system for identifying genes and determining their functions. Here we report the analysis of the genomic sequence of Arabidopsis. The sequenced regions cover 115.4 megabases of the 125-megabase genome and extend into centromeric regions. The evolution of Arabidopsis involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications, giving rise to a dynamic genome enriched by lateral gene transfer from a cyanobacterial-like ancestor of the plastid. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of Drosophila and Caenorhabditis elegans--the other sequenced multicellular eukaryotes. Arabidopsis has many families of new proteins but also lacks several common protein families, indicating that the sets of common proteins have undergone differential expansion and contraction in the three multicellular eukaryotes. This is the first complete genome sequence of a plant and provides the foundations for more comprehensive comparison of conserved processes in all eukaryotes, identifying a wide range of plant-specific gene functions and establishing rapid systematic ways to identify genes for crop improvement.

- Venter et al. (2001) — the Celera human genome.

The Sequence of the Human Genome. Venter et al. (2001). *Science*. doi:10.1126/science.1058040

> A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies-a whole-genome assembly and a regional chromosome assembly-were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional approximately 12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

- International Human Genome Sequencing Consortium (2004) — finishing the human genome.

Finishing the euchromatic sequence of the human genome. International Human Genome Sequencing Consortium (2004). *Nature*. doi:10.1038/nature03001

> The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly

> complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers approximately 99% of the euchromatic genome and is accurate to an error rate of approximately 1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000-25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

- Mouse Genome Sequencing Consortium (2002) — the mouse genome and the first comparison of two mammals.

  Initial sequencing and comparative analysis of the mouse genome. Mouse Genome Sequencing Consortium (2002). *Nature*. doi:10.1038/nature01262

  > The sequence of the mouse genome is a key informational tool for understanding the contents of the human genome and a key experimental tool for biomedical research. Here, we report the results of an international collaboration to produce a high-quality draft sequence of the mouse genome. We also present an initial comparative analysis of the mouse and human genomes, describing some of the insights that can be gleaned from the two sequences. We discuss topics including the analysis of the evolutionary forces shaping the size, structure and sequence of the genomes; the conservation of large-scale synteny across most of the genomes; the much lower extent of sequence orthology covering less than half of the genomes; the proportions of the genomes under selection; the number of protein-coding genes; the expansion of gene families related to reproduction and immunity; the evolution of proteins; and the identification of intraspecies polymorphism.

- Nurk et al. (2022) — the complete human genome.

  The complete sequence of a human genome. Nurk et al. (2022). *Science*. doi:10.1126/science.abj6987

  > Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion-base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

B.6.2 Genome size and organisation

- Thomas (1971), Gregory (2001) — the C-value paradox and the C-value enigma.

  THE GENETIC ORGANIZATION OF CHROMOSOMES. Thomas (1971). *Annual Review of Genetics*. doi:10.1146/annurev.ge.05.120171.001321

  Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Gregory (2001). *Biological Reviews of the Cambridge Philosophical Society*. doi:10.1017/s1464793100005595

  > Variation in DNA content has been largely ignored as a factor in evolution, particularly following the advent of sequence-based approaches to genomic analysis. The significant genome size diversity among organisms (more than 200000-fold among eukaryotes) bears no relationship to organismal complexity and both the origins and reasons for the clearly non-random distribution of this variation remain unclear. Several theories have been proposed to explain this 'C-value enigma' (heretofore known as the 'C-value paradox'), each of which can be described as either a mutation pressure' or 'optimal DNA' theory. Mutation pressure theories consider the large portion of non-coding DNA in eukaryotic genomes as either 'junk' or 'selfish' DNA and are important primarily in considerations of the origin of secondary DNA. Optimal DNA theories differ from mutation pressure theories by emphasizing the strong link between DNA content and cell and nuclear volumes. While mutation pressure theories generally explain this association with cell

> size as coincidental, the nucleoskeletal theory proposes a coevolutionary interaction between nuclear and cell volume, with DNA content adjusted adaptively following shifts in cell size. Each of these approaches to the C-value enigma is problematic for a variety of reasons and the preponderance of the available evidence instead favours the nucleotypic theory which postulates a causal link between bulk DNA amount and cell volume. Under this view, variation in DNA content is under direct selection via its impacts on cellular and organismal parameters. Until now, no satisfactory mechanism has been presented to explain this nucleotypic effect. However, recent advances in the study of cell cycle regulation suggest a possible 'gene nucleus interaction model' which may account for it. The present article provides a detailed review of the debate surrounding the C-value enigma, the various theories proposed to explain it, and the evidence in favour of a causal connection between DNA content and cell size. In addition, a new model of nucleotypic influence is developed, along with suggestions for further empirical investigation. Finally, some evolutionary implications of genome size diversity are considered, and a broadening of the traditional 'biological hierarchy' is recommended.

- Kapusta et al. (2017) — the dynamics of genome size: gains and losses.

  Dynamics of genome size evolution in birds and mammals. Kapusta et al. (2017). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1616702114

  > Genome size in mammals and birds shows remarkably little interspecific variation compared with other taxa. However, genome sequencing has revealed that many mammal and bird lineages have experienced differential rates of transposable element (TE) accumulation, which would be predicted to cause substantial variation in genome size between species. Thus, we hypothesize that there has been covariation between the amount of DNA gained by transposition and lost by deletion during mammal and avian evolution, resulting in genome size equilibrium. To test this model, we develop computational methods to quantify the amount of DNA gained by TE expansion and lost by deletion over the last 100 My in the lineages of 10 species of eutherian mammals and 24 species of birds. The results reveal extensive variation in the amount of DNA gained via lineage-specific transposition, but that DNA loss counteracted this expansion to various extents across lineages. Our analysis of the rate and size spectrum of deletion events implies that DNA removal in both mammals and birds has proceeded mostly through large segmental deletions (>10 kb). These findings support a unified ”accordion” model of genome size evolution in eukaryotes whereby DNA loss counteracting TE expansion is a major determinant of genome size. Furthermore, we propose that extensive DNA loss, and not necessarily a dearth of TE activity, has been the primary force maintaining the greater genomic compaction of flying birds and bats relative to their flightless relatives.

- Fernández et al. (2024) — the largest known genome.

  A 160 Gbp fork fern genome shatters size record for eukaryotes. Fernández et al. (2024). *iScience*. doi:10.1016/j.isci.2024.109889

  > Vascular plants are exceptional among eukaryotes due to their outstanding genome size diversity which ranges ~2,400-fold, including the largest genome so far recorded in the angiosperm Paris japonica (148.89 Gbp/1C). Despite available data showing that giant genomes are restricted across the Tree of Life, the biological limits to genome size expansion remain to be established. Here, we report the discovery of an even larger eukaryotic genome in Tmesipteris oblanceolata , a New Caledonian fork fern. At 160.45 Gbp/1C, this record-breaking genome challenges current understanding and opens new avenues to explore the evolutionary dynamics of genomic gigantism.

- Nakabachi et al. (2006), Giovannoni et al. (2005) — the smallest genomes: endosymbionts and streamlined bacteria.

  The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella. Nakabachi et al. (2006). *Science*. doi:10.1126/science.1134196

  > Previous studies have suggested that the minimal cellular genome could be as small as 400 kilobases. Here, we report the complete genome sequence of the psyllid symbiont Carsonella ruddii, which consists of a circular chromosome of 159,662 base pairs, averaging 16.5% GC content. It is by far the smallest and most AT-rich bacterial genome yet characterized. The genome has a high coding density (97%) with many overlapping genes and reduced gene length. Genes for translation and amino acid biosynthesis are relatively well represented, but numerous genes considered essential for life are missing, suggesting that Carsonella may have achieved organelle-like status.

Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Giovannoni et al. (2005). *Science*. doi:10.1126/science.1114057

> The SAR11 clade consists of very small, heterotrophic marine alpha-proteobacteria that are found throughout the oceans, where they account for about 25% of all microbial cells. Pelagibacter ubique, the first cultured member of this clade, has the smallest genome and encodes the smallest number of predicted open reading frames known for a free-living microorganism. In contrast to parasitic bacteria and archaea with small genomes, P. ubique has complete biosynthetic pathways for all 20 amino acids and all but a few cofactors. P. ubique has no pseudogenes, introns, transposons, extrachromosomal elements, or inteins; few paralogs; and the shortest intergenic spacers yet observed for any cell.

- Rocha (2008) — the organisation of the bacterial genome.

The Organization of the Bacterial Genome. Rocha (2008). *Annual Review of Genetics*. doi:10.1146/annurev.genet.42.110807.091653

> Many bacterial cellular processes interact intimately with the chromosome. Such interplay is the major driving force of genome structure or organization. Interactions take place at different scales-local for gene expression, global for replication-and lead to the differentiation of the chromosome into organizational units such as operons, replichores, or macrodomains. These processes are intermingled in the cell and create complex higher-level organizational features that are adaptive because they favor the interplay between the processes. The surprising result of selection for genome organization is that gene repertoires change much more quickly than chromosomal structure. Comparative genomics and experimental genomic manipulations are untangling the different cellular and evolutionary mechanisms causing such resilience to change. Since organization results from cellular processes, a better understanding of chromosome organization will help unravel the underlying cellular processes and their diversity.

- Lobry (1996) — strand asymmetry and GC skew.

Asymmetric substitution patterns in the two DNA strands of bacteria. Lobry (1996). *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a025626

> Analyses of the genomes of three prokaryotes, Escherichia coli, Bacillus subtilis, and Haemophilus influenzae, revealed a new type of genomic compartmentalization of base frequencies. There was a departure from intrastrand equifrequency between A and T or between C and G, showing that the substitution patterns of the two strands of DNA were asymmetric. The positions of the boundaries between these compartments were found to coincide with the origin and terminus of chromosome replication, and there were more A-T and C-G deviations in intergenic regions and third codon positions, suggesting that a mutational bias was responsible for this asymmetry. The strand asymmetry was found to be due to a difference in base compositions of transcripts in the leading and lagging strands. This difference is sufficient to affect codon usage, but it is small compared to the effects of gene expressivity and amino-acid composition.

- Jacob and Monod (1961) — the operon.

Genetic regulatory mechanisms in the synthesis of proteins. Jacob & Monod (1961). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(61)80072-7

- Berget et al. (1977), Chow et al. (1977) — the discovery of split genes.

Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Berget et al. (1977). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.74.8.3171

> An mRNA fraction coding for hexon polypeptide, the major virion structural protein, was purified by gel electrophoresis from extracts of adenovirus 2-infected cells late in the lytic cycle. The mRNA sequences in this fraction were mapped between 51.7 and 61.3 units on the genome by visualizing RNA-DNA hybrids in the electron microscope. When hybrids of hexon mRNA and single-stranded restriction endonuclease cleavage fragments of viral DNA were visualized in the electron microscope,branched forms were observed in which 160 nucleotides of RNA from the 5' terminus were not hydrogen bonded to the single-stranded DNA. DNA sequences complementary to the RNA sequences in each 5' tail were found by electron microscopy to be located at 17, 20, and 27 units on the same strand as that coding for the body of the hexon mRNA. Thus, four segments of viral RNA may be joined together during the synthesis of mature hexon mRNA. A model is presented for adenovirus late mRNA synthesis that involves multiple splicing during maturation of a larger precursor nuclear RNA.

An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Chow et al. (1977). *Cell.* doi:10.1016/0092-8674(77)90180-5

> The 5' terminal sequences of several adenovirus 2 (Ad2) mRNAs, isolated late in infection, are complementary to sequences within the Ad2 genome which are remote from the DNA from which the main coding sequence of each mRNA is transcribed. This has been observed by forming RNA displacement loops (R loops) between Ad2 DNA and unfractionated polysomal RNA from infected cells. The 5' terminal sequences of mRNAs in R loops, variously located between positions 36 and 92, form complex secondary hybrids with single-stranded DNA from restriction endonuclease fragments containing sequences to the left of position 36 on the Ad2 genome. The structures visualized in the electron microscope show that short sequences coded at map positions 16.6, 19.6 and 26.6 on the R strand are joined to form a leader sequence of 150-200 nucleotides at the 5' end of many late mRNAs. A late mRNA which maps to the left of position 16.6 shows a different pattern of second site hybridization. It contains sequences from 4.9-6.0 linked directly to those from 9.6-10.9. These findings imply a new mechanism for the biosynthesis of Ad2 mRNA in mammalian cells.

- Greider and Blackburn (1985) — telomerase.

Identification of a specific telomere terminal transferase activity in tetrahymena extracts. Greider & Blackburn (1985). *Cell.* doi:10.1016/0092-8674(85)90170-9

> We have found a novel activity in Tetrahymena cell free extracts that adds tandem TTGGGG repeats onto synthetic telomere primers. The single-stranded DNA oligonucleotides (TTGGGG)4 and TGTGTGGGTGTGTGGGTGTGTGGG, consisting of the Tetrahymena and yeast telomeric sequences respectively, each functioned as primers for elongation, while (CCCCAA)4 and two nontelomeric sequence DNA oligomers did not. Efficient synthesis of the TTGGGG repeats depended only on addition of micromolar concentrations of oligomer primer, dGTP, and dTTP to the extract. The activity was sensitive to heat and proteinase K treatment. The repeat addition was independent of both endogenous Tetrahymena DNA and the endogenous alpha-type DNA polymerase; and a greater elongation activity was present during macronuclear development, when a large number of telomeres are formed and replicated, than during vegetative cell growth. We propose that the novel telomere terminal transferase is involved in the addition of telomeric repeats necessary for the replication of chromosome ends in eukaryotes.

- Pan et al. (2008), Wang et al. (2008) — alternative splicing in the human transcriptome.

Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Pan et al. (2008). *Nature Genetics.* doi:10.1038/ng.259

> We carried out the first analysis of alternative splicing complexity in human tissues using mRNA-Seq data. New splice junctions were detected in approximately 20% of multiexon genes, many of which are tissue specific. By combining mRNA-Seq and EST-cDNA sequence data, we estimate that transcripts from approximately 95% of multiexon genes undergo alternative splicing and that there are approximately 100,000 intermediate- to high-abundance alternative splicing events in major human tissues. From a comparison with quantitative alternative splicing microarray profiling data, we also show that mRNA-Seq data provide reliable measurements for exon inclusion levels.

Alternative isoform regulation in human tissue transcriptomes. Wang et al. (2008). *Nature.* doi:10.1038/nature07509

> Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon-exon junctions indicated that 92-94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3' untranslated

regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

- Koonin and Wolf (2008) — the dynamic view of prokaryotic genomes.

  Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Koonin & Wolf (2008). *Nucleic Acids Research*. doi:10.1093/nar/gkn668

  > The first bacterial genome was sequenced in 1995, and the first archaeal genome in 1996. Soon after these breakthroughs, an exponential rate of genome sequencing was established, with a doubling time of approximately 20 months for bacteria and approximately 34 months for archaea. Comparative analysis of the hundreds of sequenced bacterial and dozens of archaeal genomes leads to several generalizations on the principles of genome organization and evolution. A crucial finding that enables functional characterization of the sequenced genomes and evolutionary reconstruction is that the majority of archaeal and bacterial genes have conserved orthologs in other, often, distant organisms. However, comparative genomics also shows that horizontal gene transfer (HGT) is a dominant force of prokaryotic evolution, along with the loss of genetic material resulting in genome contraction. A crucial component of the prokaryotic world is the mobilome, the enormous collection of viruses, plasmids and other selfish elements, which are in constant exchange with more stable chromosomes and serve as HGT vehicles. Thus, the prokaryotic genome space is a tightly connected, although compartmentalized, network, a novel notion that undermines the 'Tree of Life' model of evolution and requires a new conceptual framework and tools for the study of prokaryotic evolution.

B.6.3 Organelles and endosymbiosis

- Sagan (1967) — the origin of mitosing cells (endosymbiotic theory).

  On the origin of mitosing cells. Sagan (1967). *Journal of Theoretical Biology*. doi:10.1016/0022-5193(67)90079-3

  > A theory of the origin of eukaryotic cells ("higher" cells which divide by classical mitosis) is presented. By hypothesis, three fundamental organelles: the mitochondria, the photosynthetic plastids and the (9+2) basal bodies of flagella were themselves once free-living (prokaryotic) cells. The evolution of photosynthesis under the anaerobic conditions of the early atmosphere to form anaerobic bacteria, photosynthetic bacteria and eventually blue-green algae (and protoplastids) is described. The subsequent evolution of aerobic metabolism in prokaryotes to form aerobic bacteria (protoflagella and protomitochondria) presumably occurred during the transition to the oxidizing atmosphere. Classical mitosis evolved in protozoan-type cells millions of years after the evolution of photosynthesis. A plausible scheme for the origin of classical mitosis in primitive amoeboflagellates is presented. During the course of the evolution of mitosis, photosynthetic plastids (themselves derived from prokaryotes) were symbiotically acquired by some of these protozoans to form the eukaryotic algae and the green plants. The cytological, biochemical and paleontological evidence for this theory is presented, along with suggestions for further possible experimental verification. The implications of this scheme for the systematics of the lower organisms is discussed.

- Anderson et al. (1981) — the human mitochondrial genome.

  Sequence and organization of the human mitochondrial genome. Anderson et al. (1981). *Nature*. doi:10.1038/290457a0

  > The complete sequence of the 16,569-base pair human mitochondrial genome is presented. The genes for the 12S and 16S rRNAs, 22 tRNAs, cytochrome c oxidase subunits I, II and III, ATPase subunit 6, cytochrome b and eight other predicted protein coding genes have been located. The sequence shows extreme economy in that the genes have none or only a few noncoding bases between them, and in many cases the termination codons are not coded in the DNA but are created post-transcriptionally by polyadenylation of the mRNAs.

- Lang et al. (1997) — the most bacteria-like mitochondrial genome.

  An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Lang et al. (1997). *Nature*. doi:10.1038/387493a0

> Mitochondria, organelles specialized in energy conservation reactions in eukaryotic cells, have evolved from eubacteria-like endosymbionts whose closest known relatives are the rickettsial group of alpha-proteobacteria. Because characterized mitochondrial genomes vary markedly in structure, it has been impossible to infer from them the initial form of the proto-mitochondrial genome. This would require the identification of minimally derived mitochondrial DNAs that better reflect the ancestral state. Here we describe such a primitive mitochondrial genome, in the freshwater protozoon Reclinomonas americana. This protist displays ultrastructural characteristics that ally it with the retortamonads, a protozoan group that lacks mitochondria. R. americana mtDNA (69,034 base pairs) contains the largest collection of genes (97) so far identified in any mtDNA, including genes for 5S ribosomal RNA, the RNA component of RNase P, and at least 18 proteins not previously known to be encoded in mitochondria. Most surprising are four genes specifying a multisubunit, eubacterial-type RNA polymerase. Features of gene content together with eubacterial characteristics of genome organization and expression not found before in mitochondrial genomes indicate that R. americana mtDNA more closely resembles the ancestral proto-mitochondrial genome than any other mtDNA investigated to date.

- Timmis et al. (2004) — endosymbiotic gene transfer.

  Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Timmis et al. (2004). *Nature Reviews Genetics*. doi:10.1038/nrg1271

- Roger et al. (2017) — the origin and diversification of mitochondria.

  The Origin and Diversification of Mitochondria. Roger et al. (2017). *Current Biology*. doi:10.1016/j.cub.2017.09.015

  > Mitochondria are best known for their role in the generation of ATP by aerobic respiration. Yet, research in the past half century has shown that they perform a much larger suite of functions and that these functions can vary substantially among diverse eukaryotic lineages. Despite this diversity, all mitochondria derive from a common ancestral organelle that originated from the integration of an endosymbiotic alphaproteobacterium into a host cell related to Asgard Archaea. The transition from endosymbiotic bacterium to permanent organelle entailed a massive number of evolutionary changes including the origins of hundreds of new genes and a protein import system, insertion of membrane transporters, integration of metabolism and reproduction, genome reduction, endosymbiotic gene transfer, lateral gene transfer and the retargeting of proteins. These changes occurred incrementally as the endosymbiont and the host became integrated. Although many insights into this transition have been gained, controversy persists regarding the nature of the original endosymbiont, its initial interactions with the host and the timing of its integration relative to the origin of other features of eukaryote cells. Since the establishment of the organelle, proteins have been gained, lost, transferred and retargeted as mitochondria have specialized into the spectrum of functional types seen across the eukaryotic tree of life.

B.6.4 Populations

- Tettelin et al. (2005), Medini et al. (2005) — the pan-genome concept.

  Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial "pan-genome". Tettelin et al. (2005). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0506758102

  > The development of efficient and inexpensive genome sequencing methods has revolutionized the study of human bacterial pathogens and improved vaccine design. Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets. We have generated the genomic sequence of six strains representing the five major disease-causing serotypes of Streptococcus agalactiae, the main cause of neonatal infection in humans. Analysis of these genomes and those available in databases showed that the S. agalactiae species can be described by a pan-genome consisting of a core genome shared by all isolates, accounting for approximately 80% of any single genome, plus a dispensable genome consisting of partially shared and strain-specific genes. Mathematical extrapolation of the data suggests that the gene reservoir available for inclusion in the S. agalactiae pan-genome is vast and that unique genes will continue to be identified even after sequencing hundreds of genomes.

The microbial pan-genome. Medini et al. (2005). *Current Opinion in Genetics & Development*. doi:10.1016/j.gde.2005.09.006

> A decade after the beginning of the genomic era, the question of how genomics can describe a bacterial species has not been fully addressed. Experimental data have shown that in some species new genes are discovered even after sequencing the genomes of several strains. Mathematical modeling predicts that new genes will be discovered even after sequencing hundreds of genomes per species. Therefore, a bacterial species can be described by its pan-genome, which is composed of a "core genome" containing genes present in all strains, and a "dispensable genome" containing genes present in two or more strains and genes unique to single strains. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome.

- Touchon et al. (2009) — the *E. coli* pan-genome.

Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. Touchon et al. (2009). *PLoS Genetics*. doi:10.1371/journal.pgen.1000344

> The Escherichia coli species represents one of the best-studied model organisms, but also encompasses a variety of commensal and pathogenic strains that diversify by high rates of genetic change. We uniformly (re-) annotated the genomes of 20 commensal and pathogenic E. coli strains and one strain of E. fergusonii (the closest E. coli related species), including seven that we sequenced to completion. Within the approximately 18,000 families of orthologous genes, we found approximately 2,000 common to all strains. Although recombination rates are much higher than mutation rates, we show, both theoretically and using phylogenetic inference, that this does not obscure the phylogenetic signal, which places the B2 phylogenetic group and one group D strain at the basal position. Based on this phylogeny, we inferred past evolutionary events of gain and loss of genes, identifying functional classes under opposite selection pressures. We found an important adaptive role for metabolism diversification within group B2 and Shigella strains, but identified few or no extraintestinal virulence-specific genes, which could render difficult the development of a vaccine against extraintestinal infections. Genome flux in E. coli is confined to a small number of conserved positions in the chromosome, which most often are not associated with integrases or tRNA genes. Core genes flanking some of these regions show higher rates of recombination, suggesting that a gene, once acquired by a strain, spreads within the species by homologous recombination at the flanking genes. Finally, the genome's long-scale structure of recombination indicates lower recombination rates, but not higher mutation rates, at the terminus of replication. The ensuing effect of background selection and biased gene conversion may thus explain why this region is A+T-rich and shows high sequence divergence but low sequence polymorphism. Overall, despite a very high gene flow, genes co-exist in an organised genome.

- Sela et al. (2016) — a theory of prokaryotic genome evolution.

Theory of prokaryotic genome evolution. Sela et al. (2016). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1614083113

> Bacteria and archaea typically possess small genomes that are tightly packed with protein-coding genes. The compactness of prokaryotic genomes is commonly perceived as evidence of adaptive genome streamlining caused by strong purifying selection in large microbial populations. In such populations, even the small cost incurred by nonfunctional DNA because of extra energy and time expenditure is thought to be sufficient for this extra genetic material to be eliminated by selection. However, contrary to the predictions of this model, there exists a consistent, positive correlation between the strength of selection at the protein sequence level, measured as the ratio of nonsynonymous to synonymous substitution rates, and microbial genome size. Here, by fitting the genome size distributions in multiple groups of prokaryotes to predictions of mathematical models of population evolution, we show that only models in which acquisition of additional genes is, on average, slightly beneficial yield a good fit to genomic data. These results suggest that the number of genes in prokaryotic genomes reflects the equilibrium between the benefit of additional genes that diminishes as the genome grows and deletion bias (i.e., the rate of deletion of genetic material being slightly greater than the rate of acquisition). Thus, new genes acquired by microbial genomes, on average, appear to be adaptive. The tight spacing of protein-coding genes likely results from a combination of the deletion bias and purifying selection that efficiently eliminates nonfunctional, noncoding sequences.

- Lynch et al. (2016) — genetic drift, selection and the evolution of the mutation rate.

Genetic drift, selection and the evolution of the mutation rate. Lynch et al. (2016). *Nature Reviews Genetics*. doi:10.1038/nrg.2016.104

> As one of the few cellular traits that can be quantified across the tree of life, DNA-replication fidelity provides an excellent platform for understanding fundamental evolutionary processes. Furthermore, because mutation is the ultimate source of all genetic variation, clarifying why mutation rates vary is crucial for understanding all areas of biology. A potentially revealing hypothesis for mutation-rate evolution is that natural selection primarily operates to improve replication fidelity, with the ultimate limits to what can be achieved set by the power of random genetic drift. This drift-barrier hypothesis is consistent with comparative measures of mutation rates, provides a simple explanation for the existence of error-prone polymerases and yields a formal counter-argument to the view that selection fine-tunes gene-specific mutation rates.

- Cann et al. (1987), Hublin et al. (2017) — mitochondrial Eve and the age of *Homo sapiens*.

Mitochondrial DNA and human evolution. Cann et al. (1987). *Nature*. doi:10.1038/325031a0

> Mitochondrial DNAs from 147 people, drawn from five geographic populations have been analysed by restriction mapping. All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa. All the populations examined except the African population have multiple origins, implying that each area was colonised repeatedly.

New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Hublin et al. (2017). *Nature*. doi:10.1038/nature22336

> Fossil evidence points to an African origin of Homo sapiens from a group called either H. heidelbergensis or H. rhodesiensis. However, the exact place and time of emergence of H. sapiens remain obscure because the fossil record is scarce and the chronological age of many key specimens remains uncertain. In particular, it is unclear whether the present day 'modern' morphology rapidly emerged approximately 200 thousand years ago (ka) among earlier representatives of H. sapiens or evolved gradually over the last 400 thousand years. Here we report newly discovered human fossils from Jebel Irhoud, Morocco, and interpret the affinities of the hominins from this site with other archaic and recent human groups. We identified a mosaic of features including facial, mandibular and dental morphology that aligns the Jebel Irhoud material with early or recent anatomically modern humans and more primitive neurocranial and endocranial morphology. In combination with an age of 315 ± 34 thousand years (as determined by thermoluminescence dating), this evidence makes Jebel Irhoud the oldest and richest African Middle Stone Age hominin site that documents early stages of the H. sapiens clade in which key features of modern morphology were established. Furthermore, it shows that the evolutionary processes behind the emergence of H. sapiens involved the whole African continent.

- Krings et al. (1997) — the first Neanderthal DNA.

Neandertal DNA Sequences and the Origin of Modern Humans. Krings et al. (1997). *Cell*. doi:10.1016/s0092-8674(00)80310-4

> DNA was extracted from the Neandertal-type specimen found in 1856 in western Germany. By sequencing clones from short overlapping PCR products, a hitherto unknown mitochondrial (mt) DNA sequence was determined. Multiple controls indicate that this sequence is endogenous to the fossil. Sequence comparisons with human mtDNA sequences, as well as phylogenetic analyses, show that the Neandertal sequence falls outside the variation of modern humans. Furthermore, the age of the common ancestor of the Neandertal and modern human mtDNAs is estimated to be four times greater than that of the common ancestor of human mtDNAs. This suggests that Neandertals went extinct without contributing mtDNA to modern humans.

- Reich et al. (2010), Meyer et al. (2012) — the Denisovans.

Genetic history of an archaic hominin group from Denisova Cave in Siberia. Reich et al. (2010). *Nature*. doi:10.1038/nature09710

> Using DNA extracted from a finger bone found in Denisova Cave in southern Siberia, we have sequenced the genome of an archaic hominin to about 1.9-fold coverage. This individual is from a group that shares a common origin with Neanderthals. This population was not involved in the putative gene flow from Neanderthals into Eurasians; however, the data suggest that it contributed 4-6% of its genetic material to the genomes of present-day Melanesians. We designate this hominin population 'Denisovans' and suggest that it may have been widespread in Asia during the Late Pleistocene epoch. A tooth found in Denisova Cave carries a mitochondrial genome highly similar to that of the finger bone. This tooth shares no derived morphological features with Neanderthals or modern humans, further indicating that Denisovans have an evolutionary history distinct from Neanderthals and modern humans.

A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Meyer et al. (2012). *Science*. doi:10.1126/science.1224344

> We present a DNA library preparation method that has allowed us to reconstruct a high-coverage (30×) genome sequence of a Denisovan, an extinct relative of Neandertals. The quality of this genome allows a direct estimation of Denisovan heterozygosity indicating that genetic diversity in these archaic hominins was extremely low. It also allows tentative dating of the specimen on the basis of "missing evolution" in its genome, detailed measurements of Denisovan and Neandertal admixture into present-day human populations, and the generation of a near-complete catalog of genetic changes that swept to high frequency in modern humans since their divergence from Denisovans.

- Prüfer et al. (2013) — a high-coverage Neanderthal genome.

The complete genome sequence of a Neanderthal from the Altai Mountains. Prüfer et al. (2013). *Nature*. doi:10.1038/nature12886

> We present a high-quality genome sequence of a Neanderthal woman from Siberia. We show that her parents were related at the level of half-siblings and that mating among close relatives was common among her recent ancestors. We also sequenced the genome of a Neanderthal from the Caucasus to low coverage. An analysis of the relationships and population history of available archaic genomes and 25 present-day human genomes shows that several gene flow events occurred among Neanderthals, Denisovans and early modern humans, possibly including gene flow into Denisovans from an unknown archaic group. Thus, interbreeding, albeit of low magnitude, occurred among many hominin groups in the Late Pleistocene. In addition, the high-quality Neanderthal genome allows us to establish a definitive list of substitutions that became fixed in modern humans after their separation from the ancestors of Neanderthals and Denisovans.

- Slon et al. (2018) — a Neanderthal–Denisovan hybrid.

The genome of the offspring of a Neanderthal mother and a Denisovan father. Slon et al. (2018). *Nature*. doi:10.1038/s41586-018-0455-x

> Neanderthals and Denisovans are extinct groups of hominins that separated from each other more than 390,000 years ago 1,2 . Here we present the genome of 'Denisova 11', a bone fragment from Denisova Cave (Russia) 3 and show that it comes from an individual who had a Neanderthal mother and a Denisovan father. The father, whose genome bears traces of Neanderthal ancestry, came from a population related to a later Denisovan found in the cave 4-6 . The mother came from a population more closely related to Neanderthals who lived later in Europe 2,7 than to an earlier Neanderthal found in Denisova Cave 8 , suggesting that migrations of Neanderthals between eastern and western Eurasia occurred sometime after 120,000 years ago. The finding of a first-generation Neanderthal-Denisovan offspring among the small number of archaic specimens sequenced to date suggests that mixing between Late Pleistocene hominin groups was common when they met.

- Sankararaman et al. (2014) — the landscape of Neanderthal ancestry in present-day humans.

The genomic landscape of Neanderthal ancestry in present-day humans. Sankararaman et al. (2014). *Nature*. doi:10.1038/nature12961

> Genomic studies have shown that Neanderthals interbred with modern humans, and that non-Africans today are the products of this mixture. The antiquity of Neanderthal gene flow into modern humans means that genomic regions that derive from Neanderthals in any one human today are usually less than a hundred kilobases in size. However, Neanderthal haplotypes are also distinctive enough that several studies have been able to detect Neanderthal ancestry at specific loci. We systematically infer Neanderthal

> haplotypes in the genomes of 1,004 present-day humans. Regions that harbour a high frequency of Neanderthal alleles are enriched for genes affecting keratin filaments, suggesting that Neanderthal alleles may have helped modern humans to adapt to non-African environments. We identify multiple Neanderthal-derived alleles that confer risk for disease, suggesting that Neanderthal alleles continue to shape human biology. An unexpected finding is that regions with reduced Neanderthal ancestry are enriched in genes, implying selection to remove genetic material derived from Neanderthals. Genes that are more highly expressed in testes than in any other tissue are especially reduced in Neanderthal ancestry, and there is an approximately fivefold reduction of Neanderthal ancestry on the X chromosome, which is known from studies of diverse species to be especially dense in male hybrid sterility genes. These results suggest that part of the explanation for genomic regions of reduced Neanderthal ancestry is Neanderthal alleles that caused decreased fertility in males when moved to a modern human genetic background.

- Huerta-Sánchez et al. (2014) — adaptive introgression in Tibetans.

  Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Huerta-Sánchez et al. (2014). *Nature*. doi:10.1038/nature13408

  > As modern humans migrated out of Africa, they encountered many new environmental conditions, including greater temperature extremes, different pathogens and higher altitudes. These diverse environments are likely to have acted as agents of natural selection and to have led to local adaptations. One of the most celebrated examples in humans is the adaptation of Tibetans to the hypoxic environment of the high-altitude Tibetan plateau. A hypoxia pathway gene, EPAS1, was previously identified as having the most extreme signature of positive selection in Tibetans, and was shown to be associated with differences in haemoglobin concentration at high altitude. Re-sequencing the region around EPAS1 in 40 Tibetan and 40 Han individuals, we find that this gene has a highly unusual haplotype structure that can only be convincingly explained by introgression of DNA from Denisovan or Denisovan-related individuals into humans. Scanning a larger set of worldwide populations, we find that the selected haplotype is only found in Denisovans and in Tibetans, and at very low frequency among Han Chinese. Furthermore, the length of the haplotype, and the fact that it is not found in any other populations, makes it unlikely that the haplotype sharing between Tibetans and Denisovans was caused by incomplete ancestral lineage sorting rather than introgression. Our findings illustrate that admixture with other hominin species has provided genetic variation that helped humans to adapt to new environments.

- The 1000 Genomes Project Consortium (2015) — the 1000 Genomes Project.

  A global reference for human genetic variation. The 1000 Genomes Project Consortium (2015). *Nature*. doi:10.1038/nature15393

  > The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

B.6.5 Repeats and the 3D genome

- McClintock (1950) — the discovery of transposable elements.

  The origin and behavior of mutable loci in maize. McClintock (1950). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.36.6.344

- Wells and Feschotte (2020) — a field guide to transposable elements.

  A Field Guide to Eukaryotic Transposable Elements. Wells & Feschotte (2020). *Annual Review of Genetics*. doi:10.1146/annurev-genet-040620-022145

> Transposable elements (TEs) are mobile DNA sequences that propagate within genomes. Through diverse invasion strategies, TEs have come to occupy a substantial fraction of nearly all eukaryotic genomes, and they represent a major source of genetic variation and novelty. Here we review the defining features of each major group of eukaryotic TEs and explore their evolutionary origins and relationships. We discuss how the unique biology of different TEs influences their propagation and distribution within and across genomes. Environmental and genetic factors acting at the level of the host species further modulate the activity, diversification, and fate of TEs, producing the dramatic variation in TE content observed across eukaryotes. We argue that cataloging TE diversity and dissecting the idiosyncratic behavior of individual elements are crucial to expanding our comprehension of their impact on the biology of genomes and the evolution of species.

- Storer et al. (2021) — Dfam.

  The Dfam community resource of transposable element families, sequence models, and genome annotations. Storer et al. (2021). *Mobile DNA*. doi:10.1186/s13100-020-00230-y

  > Dfam is an open access database of repetitive DNA families, sequence models, and genome annotations. The 3.0-3.3 releases of Dfam (https://dfam.org) represent an evolution from a proof-of-principle collection of transposable element families in model organisms into a community resource for a broad range of species, and for both curated and uncurated datasets. In addition, releases since Dfam 3.0 provide auxiliary consensus sequence models, transposable element protein alignments, and a formalized classification system to support the growing diversity of organisms represented in the resource. The latest release includes 266,740 new de novo generated transposable element families from 336 species contributed by the EBI. This expansion demonstrates the utility of many of Dfam's new features and provides insight into the long term challenges ahead for improving de novo generated transposable element datasets.

- Bailey et al. (2002) — segmental duplications in the human genome.

  Recent Segmental Duplications in the Human Genome. Bailey et al. (2002). *Science*. doi:10.1126/science.1072047

  > Primate-specific segmental duplications are considered important in human disease and evolution. The inability to distinguish between allelic and duplication sequence overlap has hampered their characterization as well as assembly and annotation of our genome. We developed a method whereby each public sequence is analyzed at the clone level for overrepresentation within a whole-genome shotgun sequence. This test has the ability to detect duplications larger than 15 kilobases irrespective of copy number, location, or high sequence similarity. We mapped 169 large regions flanked by highly similar duplications. Twenty-four of these hot spots of genomic instability have been associated with genetic disease. Our analysis indicates a highly nonrandom chromosomal and genic distribution of recent segmental duplications, with a likely role in expanding protein diversity.

- Altemose et al. (2022) — complete maps of human centromeres.

  Complete genomic and epigenetic maps of human centromeres. Altemose et al. (2022). *Science*. doi:10.1126/science.abl4178

  > Existing human genome assemblies have almost entirely excluded repetitive sequences within and near centromeres, limiting our understanding of their organization, evolution, and functions, which include facilitating proper chromosome segregation. Now, a complete, telomere-to-telomere human genome assembly (T2T-CHM13) has enabled us to comprehensively characterize pericentromeric and centromeric repeats, which constitute 6.2% of the genome (189.9 megabases). Detailed maps of these regions revealed multimegabase structural rearrangements, including in active centromeric repeat arrays. Analysis of centromere-associated sequences uncovered a strong relationship between the position of the centromere and the evolution of the surrounding DNA through layered repeat expansions. Furthermore, comparisons of chromosome X centromeres across a diverse panel of individuals illuminated high degrees of structural, epigenetic, and sequence variation in these complex and rapidly evolving regions.

- Dekker et al. (2002) — chromosome conformation capture.

  Capturing Chromosome Conformation. Dekker et al. (2002). *Science*. doi:10.1126/science.1067799

> We describe an approach to detect the frequency of interaction between any two genomic loci. Generation of a matrix of interaction frequencies between sites on the same or different chromosomes reveals their relative spatial disposition and provides information about the physical properties of the chromatin fiber. This methodology can be applied to the spatial organization of entire genomes in organisms from bacteria to human. Using the yeast Saccharomyces cerevisiae, we could confirm known qualitative features of chromosome organization within the nucleus and dynamic changes in that organization during meiosis. We also analyzed yeast chromosome III at the G1 stage of the cell cycle. We found that chromatin is highly flexible throughout. Furthermore, functionally distinct AT- and GC-rich domains were found to exhibit different conformations, and a population-average 3D model of chromosome III could be determined. Chromosome III emerges as a contorted ring.

- Lieberman-Aiden et al. (2009) — Hi-C.

  Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Lieberman-Aiden et al. (2009). *Science*. doi:10.1126/science.1181369

  > We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase. These maps confirm the presence of chromosome territories and the spatial proximity of small, gene-rich chromosomes. We identified an additional level of genome organization that is characterized by the spatial segregation of open and closed chromatin to form two genome-wide compartments. At the megabase scale, the chromatin conformation is consistent with a fractal globule, a knot-free, polymer conformation that enables maximally dense packing while preserving the ability to easily fold and unfold any genomic locus. The fractal globule is distinct from the more commonly used globular equilibrium model. Our results demonstrate the power of Hi-C to map the dynamic conformations of whole genomes.

- Dixon et al. (2012), Nora et al. (2012) — topologically associating domains.

  Topological domains in mammalian genomes identified by analysis of chromatin interactions. Dixon et al. (2012). *Nature*. doi:10.1038/nature11082

  > The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is coarse, fragmented and incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories, and numerous models have been proposed for how chromosomes fold within chromosome territories. These models, however, provide only few mechanistic details about the relationship between higher order chromatin structure and genome function. Recent advances in genomic technologies have led to rapid advances in the study of three-dimensional genome organization. In particular, Hi-C has been introduced as a method for identifying higher order chromatin interactions genome wide. Here we investigate the three-dimensional organization of the human and mouse genomes in embryonic stem cells and terminally differentiated cell types at unprecedented resolution. We identify large, megabase-sized local chromatin interaction domains, which we term 'topological domains', as a pervasive structural feature of the genome organization. These domains correlate with regions of the genome that constrain the spread of heterochromatin. The domains are stable across different cell types and highly conserved across species, indicating that topological domains are an inherent property of mammalian genomes. Finally, we find that the boundaries of topological domains are enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons, indicating that these factors may have a role in establishing the topological domain structure of the genome.

  Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nora et al. (2012). *Nature*. doi:10.1038/nature11049

  > In eukaryotes transcriptional regulation often involves multiple long-range elements and is influenced by the genomic environment. A prime example of this concerns the mouse X-inactivation centre (Xic), which orchestrates the initiation of X-chromosome inactivation (XCI) by controlling the expression of the non-protein-coding Xist transcript. The extent of Xic sequences required for the proper regulation of Xist remains unknown. Here we use chromosome conformation capture carbon-copy (5C) and super-resolution microscopy to analyse the spatial organization of a 4.5-megabases (Mb) region including Xist. We discover a series of discrete 200-kilobase to 1 Mb topologically associating domains (TADs), present both before and after cell differentiation and on the active and inactive X. TADs align with, but do not rely on, several domain-wide features of the epigenome, such as H3K27me3 or H3K9me2 blocks and lamina-associated domains. TADs also align with coordinately regulated gene clusters. Disruption of a

TAD boundary causes ectopic chromosomal contacts and long-range transcriptional misregulation. The Xist/Tsix sense/antisense unit illustrates how TADs enable the spatial segregation of oppositely regulated chromosomal neighbourhoods, with the respective promoters of Xist and Tsix lying in adjacent TADs, each containing their known positive regulators. We identify a novel distal regulatory region of Tsix within its TAD, which produces a long intervening RNA, Linx. In addition to uncovering a new principle of cis-regulatory architecture of mammalian chromosomes, our study sets the stage for the full genetic dissection of the X-inactivation centre.

- Rao et al. (2014), Fudenberg et al. (2016) — loops and loop extrusion.

A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Rao et al. (2014). *Cell*. doi:10.1016/j.cell.2014.11.021

We use in situ Hi-C to probe the 3D architecture of genomes, constructing haploid and diploid maps of nine cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1 kb resolution. We find that genomes are partitioned into contact domains (median length, 185 kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs "facing" one another. The inactive X chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats.

Formation of Chromosomal Domains by Loop Extrusion. Fudenberg et al. (2016). *Cell Reports*. doi:10.1016/j.celrep.2016.04.085

Topologically associating domains (TADs) are fundamental structural and functional building blocks of human interphase chromosomes, yet the mechanisms of TAD formation remain unclear. Here, we propose that loop extrusion underlies TAD formation. In this process, cis-acting loop-extruding factors, likely cohesins, form progressively larger loops but stall at TAD boundaries due to interactions with boundary proteins, including CTCF. Using polymer simulations, we show that this model produces TADs and finer-scale features of Hi-C data. Each TAD emerges from multiple loops dynamically formed through extrusion, contrary to typical illustrations of single static loops. Loop extrusion both explains diverse experimental observations-including the preferential orientation of CTCF motifs, enrichments of architectural proteins at TAD boundaries, and boundary deletion experiments-and makes specific predictions for the depletion of CTCF versus cohesin. Finally, loop extrusion has potentially far-ranging consequences for processes such as enhancer-promoter interactions, orientation-specific chromosomal looping, and compaction of mitotic chromosomes.

B.6.6 Non-coding RNA and the function debate

- Lee et al. (1993), Bartel (2018) — microRNAs.

The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Lee et al. (1993). *Cell*. doi:10.1016/0092-8674(93)90529-y

lin-4 is essential for the normal temporal control of diverse postembryonic developmental events in C. elegans. lin-4 acts by negatively regulating the level of LIN-14 protein, creating a temporal decrease in LIN-14 protein starting in the first larval stage (L1). We have cloned the C. elegans lin-4 locus by chromosomal walking and transformation rescue. We used the C. elegans clone to isolate the gene from three other Caenorhabditis species; all four Caenorhabditis clones functionally rescue the lin-4 null allele of C. elegans. Comparison of the lin-4 genomic sequence from these four species and site-directed mutagenesis of potential open reading frames indicated that lin-4 does not encode a protein. Two small lin-4 transcripts of approximately 22 and 61 nt were identified in C. elegans and found to contain sequences complementary to a repeated sequence element in the 3' untranslated region (UTR) of lin-14 mRNA, suggesting that lin-4 regulates lin-14 translation via an antisense RNA-RNA interaction.

Metazoan MicroRNAs. Bartel (2018). *Cell*. doi:10.1016/j.cell.2018.03.006

> MicroRNAs (miRNAs) are ~22 nt RNAs that direct posttranscriptional repression of mRNA targets in diverse eukaryotic lineages. In humans and other mammals, these small RNAs help sculpt the expression of most mRNAs. This article reviews advances in our understanding of the defining features of metazoan miRNAs and their biogenesis, genomics, and evolution. It then reviews how metazoan miRNAs are regulated, how they recognize and cause repression of their targets, and the biological functions of this repression, with a compilation of knockout phenotypes that shows that important biological functions have been identified for most of the broadly conserved miRNAs of mammals.

- Mattick et al. (2023) — long non-coding RNAs.

  Long non-coding RNAs: definitions, functions, challenges and recommendations. Mattick et al. (2023). *Nature Reviews Molecular Cell Biology*. doi:10.1038/s41580-022-00566-8

  > Genes specifying long non-coding RNAs (lncRNAs) occupy a large fraction of the genomes of complex organisms. The term 'lncRNAs' encompasses RNA polymerase I (Pol I), Pol II and Pol III transcribed RNAs, and RNAs from processed introns. The various functions of lncRNAs and their many isoforms and interleaved relationships with other genes make lncRNA classification and annotation difficult. Most lncRNAs evolve more rapidly than protein-coding sequences, are cell type specific and regulate many aspects of cell differentiation and development and other physiological processes. Many lncRNAs associate with chromatin-modifying complexes, are transcribed from enhancers and nucleate phase separation of nuclear condensates and domains, indicating an intimate link between lncRNA expression and the spatial control of gene expression during development. lncRNAs also have important roles in the cytoplasm and beyond, including in the regulation of translation, metabolism and signalling. lncRNAs often have a modular structure and are rich in repeats, which are increasingly being shown to be relevant to their function. In this Consensus Statement, we address the definition and nomenclature of lncRNAs and their conservation, expression, phenotypic visibility, structure and functions. We also discuss research challenges and provide recommendations to advance the understanding of the roles of lncRNAs in development, cell biology and disease.

- The ENCODE Project Consortium (2012), The ENCODE Project Consortium (2020) — ENCODE.

  An integrated encyclopedia of DNA elements in the human genome. The ENCODE Project Consortium (2012). *Nature*. doi:10.1038/nature11247

  > The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

  Expanded encyclopaedias of DNA elements in the human and mouse genomes. The ENCODE Project Consortium (2020). *Nature*. doi:10.1038/s41586-020-2493-4

  > The human and mouse genomes contain instructions that specify RNAs and proteins and govern the timing, magnitude, and cellular context of their production. To better delineate these elements, phase III of the Encyclopedia of DNA Elements (ENCODE) Project has expanded analysis of the cell and tissue repertoires of RNA transcription, chromatin structure and modification, DNA methylation, chromatin looping, and occupancy by transcription factors and RNA-binding proteins. Here we summarize these efforts, which have produced 5,992 new experimental datasets, including systematic determinations across mouse fetal development. All data are available through the ENCODE data portal (https://www.encodeproject.org), including phase II ENCODE 1 and Roadmap Epigenomics 2 data. We have developed a registry of 926,535 human and 339,815 mouse candidate cis-regulatory elements, covering 7.9 and 3.4% of their respective genomes, by integrating selected datatypes associated with gene regulation, and constructed a web-based server (SCREEN; http://screen.encodeproject.org) to provide flexible, user-defined access to this resource. Collectively, the ENCODE data and registry provide an expansive resource for the scientific community to build a better understanding of the organization and function of the human and mouse genomes.

- Graur et al. (2013), Doolittle (2013) — critiques of ENCODE.

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. Graur et al. (2013). *Genome Biology and Evolution*. doi:10.1093/gbe/evt028

> A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least 80 - 10 = 70% of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly by employing the seldom used "causal role" definition of biological function and then applying it inconsistently to different biochemical properties, by committing a logical fallacy known as "affirming the consequent," by failing to appreciate the crucial difference between "junk DNA" and "garbage DNA," by using analytical methods that yield biased errors and inflate estimates of functionality, by favoring statistical sensitivity over specificity, and by emphasizing statistical significance rather than the magnitude of the effect. Here, we detail the many logical and methodological transgressions involved in assigning functionality to almost every nucleotide in the human genome. The ENCODE results were predicted by one of its authors to necessitate the rewriting of textbooks. We agree, many textbooks dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

Is junk DNA bunk? A critique of ENCODE. Doolittle (2013). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1221376110

> Do data from the Encyclopedia Of DNA Elements (ENCODE) project render the notion of junk DNA obsolete? Here, I review older arguments for junk grounded in the C-value paradox and propose a thought experiment to challenge ENCODE's ontology. Specifically, what would we expect for the number of functional elements (as ENCODE defines them) in genomes much larger than our own genome? If the number were to stay more or less constant, it would seem sensible to consider the rest of the DNA of larger genomes to be junk or, at least, assign it a different sort of role (structural rather than informational). If, however, the number of functional elements were to rise significantly with C-value then, (i) organisms with genomes larger than our genome are more complex phenotypically than we are, (ii) ENCODE's definition of functional element identifies many sites that would not be considered functional or phenotype-determining by standard uses in biology, or (iii) the same phenotypic functions are often determined in a more diffuse fashion in larger-genomed organisms. Good cases can be made for propositions ii and iii. A larger theoretical framework, embracing informational and structural roles for DNA, neutral as well as adaptive causes of complexity, and selection as a multilevel phenomenon, is needed.

- Kellis et al. (2014) — defining functional DNA elements.

Defining functional DNA elements in the human genome. Kellis et al. (2014). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1318948111

> With the completion of the human genome sequence, attention turned to identifying and annotating its functional DNA elements. As a complement to genetic and comparative genomics approaches, the Encyclopedia of DNA Elements Project was launched to contribute maps of RNA transcripts, transcriptional regulator binding sites, and chromatin states in many cell types. The resulting genome-wide data reveal sites of biochemical activity with high positional resolution and cell type specificity that facilitate studies of gene regulation and interpretation of noncoding variants associated with human disease. However, the biochemically active regions cover a much larger fraction of the genome than do evolutionarily conserved regions, raising the question of whether nonconserved but biochemically active regions are truly functional. Here, we review the strengths and limitations of biochemical, evolutionary, and genetic approaches for defining functional DNA segments, potential sources for the observed differences in estimated genomic coverage, and the biological implications of these discrepancies. We also analyze the relationship between signal intensity, genomic coverage, and evolutionary conservation. Our results reinforce the principle that each approach provides complementary information and that we need to use combinations of all three to elucidate genome function in human biology and disease.

- Rands et al. (2014), Christmas et al. (2023) — the constrained fraction of the human genome.

  8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. Rands et al. (2014). *PLoS Genetics*. doi:10.1371/journal.pgen.1004525

  > Ten years on from the finishing of the human reference genome sequence, it remains unclear what fraction of the human genome confers function, where this sequence resides, and how much is shared with other mammalian species. When addressing these questions, functional sequence has often been equated with pan-mammalian conserved sequence. However, functional elements that are short-lived, including those contributing to species-specific biology, will not leave a footprint of long-lasting negative selection. Here, we address these issues by identifying and characterising sequence that has been constrained with respect to insertions and deletions for pairs of eutherian genomes over a range of divergences. Within noncoding sequence, we find increasing amounts of mutually constrained sequence as species pairs become more closely related, indicating that noncoding constrained sequence turns over rapidly. We estimate that half of present-day noncoding constrained sequence has been gained or lost in approximately the last 130 million years (half-life in units of divergence time, d1/2 = 0.25-0.31). While enriched with ENCODE biochemical annotations, much of the short-lived constrained sequences we identify are not detected by models optimized for wider pan-mammalian conservation. Constrained DNase 1 hypersensitivity sites, promoters and untranslated regions have been more evolutionarily stable than long noncoding RNA loci which have turned over especially rapidly. By contrast, protein coding sequence has been highly stable, with an estimated half-life of over a billion years (d1/2 = 2.1-5.0). From extrapolations we estimate that 8.2% (7.1-9.2%) of the human genome is presently subject to negative selection and thus is likely to be functional, while only 2.2% has maintained constraint in both human and mouse since these species diverged. These results reveal that the evolutionary history of the human genome has been highly dynamic, particularly for its noncoding yet biologically functional fraction.

  Evolutionary constraint and innovation across hundreds of placental mammals. Christmas et al. (2023). *Science*. doi:10.1126/science.abn3943

  > Zoonomia is the largest comparative genomics resource for mammals produced to date. By aligning genomes for 240 species, we identify bases that, when mutated, are likely to affect fitness and alter disease risk. At least 332 million bases (~10.7%) in the human genome are unusually conserved across species (evolutionarily constrained) relative to neutrally evolving repeats, and 4552 ultraconserved elements are nearly perfectly conserved. Of 101 million significantly constrained single bases, 80% are outside protein-coding exons and half have no functional annotations in the Encyclopedia of DNA Elements (ENCODE) resource. Changes in genes and regulatory elements are associated with exceptional mammalian traits, such as hibernation, that could inform therapeutic development. Earth's vast and imperiled biodiversity offers distinctive power for identifying genetic variants that affect genome function and organismal phenotypes.

- Koonin (2011) — are there laws of genome evolution?

  Are There Laws of Genome Evolution?. Koonin (2011). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002173

  > Research in quantitative evolutionary genomics and systems biology led to the discovery of several universal regularities connecting genomic and molecular phenomic variables. These universals include the log-normal distribution of the evolutionary rates of orthologous genes; the power law-like distributions of paralogous family size and node degree in various biological networks; the negative correlation between a gene's sequence evolution rate and expression level; and differential scaling of functional classes of genes with genome size. The universals of genome evolution can be accounted for by simple mathematical models similar to those used in statistical physics, such as the birth-death-innovation model. These models do not explicitly incorporate selection; therefore, the observed universal regularities do not appear to be shaped by selection but rather are emergent properties of gene ensembles. Although a complete physical theory of evolutionary biology is inconceivable, the universals of genome evolution might qualify as "laws of evolutionary genomics" in the same sense "law" is understood in modern physics.

B.7 Expression analysis

Further papers for Chapter 7.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.7.1 Microarrays

- Velculescu et al. (1995) — serial analysis of gene expression (SAGE).

  Serial Analysis of Gene Expression. Velculescu et al. (1995). *Science*. doi:10.1126/science.270.5235.484

  > The characteristics of an organism are determined by the genes expressed within it. A method was developed, called serial analysis of gene expression (SAGE), that allows the quantitative and simultaneous analysis of a large number of transcripts. To demonstrate this strategy, short diagnostic sequence tags were isolated from pancreas, concatenated, and cloned. Manual sequencing of 1000 tags revealed a gene expression pattern characteristic of pancreatic function. New pancreatic transcripts corresponding to novel tags were identified. SAGE should provide a broadly applicable means for the quantitative cataloging and comparison of expressed genes in a variety of normal, developmental, and disease states.

- Fodor et al. (1991), Lockhart et al. (1996) — Affymetrix oligonucleotide arrays.

  Light-Directed, Spatially Addressable Parallel Chemical Synthesis. Fodor et al. (1991). *Science*. doi:10.1126/science.1990438

  > Solid-phase chemistry, photolabile protecting groups, and photolithography have been combined to achieve light-directed, spatially addressable parallel chemical synthesis to yield a highly diverse set of chemical products. Binary masking, one of many possible combinatorial synthesis strategies, yields 2n compounds in n chemical steps. An array of 1024 peptides was synthesized in ten steps, and its interaction with a monoclonal antibody was assayed by epifluorescence microscopy. High-density arrays formed by light-directed synthesis are potentially rich sources of chemical diversity for discovering new ligands that bind to biological receptors and for elucidating principles governing molecular interactions. The generality of this approach is illustrated by the light-directed synthesis of a dinucleotide. Spatially directed synthesis of complex compounds could also be used for microfabrication of devices.

  Expression monitoring by hybridization to high-density oligonucleotide arrays. Lockhart et al. (1996). *Nature Biotechnology*. doi:10.1038/nbt1296-1675

  > The human genome encodes approximately 100,000 different genes, and at least partial sequence information for nearly all will be available soon. Sequence information alone, however, is insufficient for a full understanding of gene function, expression, regulation, and splice-site variation. Because cellular processes are governed by the repertoire of expressed genes, and the levels and timing of expression, it is important to have experimental tools for the direct monitoring of large numbers of mRNAs in parallel. We have developed an approach that is based on hybridization to small, high-density arrays containing tens of thousands of synthetic oligonucleotides. The arrays are designed based on sequence information alone and are synthesized in situ using a combination of photolithography and oligonucleotide chemistry. RNAs present at a frequency of 1:300,000 are unambiguously detected, and detection is quantitative over more than three orders of magnitude. This approach provides a way to use directly the growing body of sequence information for highly parallel experimental investigations. Because of the combinatorial nature of the chemistry and the ability to synthesize small arrays containing hundreds of thousands of specifically chosen oligonucleotides, the method is readily scalable to the simultaneous monitoring of tens of thousands of genes.

- Schena et al. (1995) — spotted cDNA microarrays.

  Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Schena et al. (1995). *Science*. doi:10.1126/science.270.5235.467

> A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

- Gunderson et al. (2004), Kuhn et al. (2004) — Illumina BeadArrays.

  Decoding Randomly Ordered DNA Arrays. Gunderson et al. (2004). *Genome Research*. doi:10.1101/gr.2255804

  > We have developed a simple and efficient algorithm to identify each member of a large collection of DNA-linked objects through the use of hybridization, and have applied it to the manufacture of randomly assembled arrays of beads in wells. Once the algorithm has been used to determine the identity of each bead, the microarray can be used in a wide variety of applications, including single nucleotide polymorphism genotyping and gene expression profiling. The algorithm requires only a few labels and several sequential hybridizations to identify thousands of different DNA sequences with great accuracy. We have decoded tens of thousands of arrays, each with 1520 sequences represented at approximately 30-fold redundancy by up to approximately 50,000 beads, with a median error rate of <1 x 10(-4) per bead. The approach makes use of error checking codes and provides, for the first time, a direct functional quality control of every element of each array that is manufactured. The algorithm can be applied to any spatially fixed collection of objects or molecules that are associated with specific DNA sequences.

  A novel, high-performance random array platform for quantitative gene expression profiling. Kuhn et al. (2004). *Genome Research*. doi:10.1101/gr.2739104

  > We have developed a new microarray technology for quantitative gene-expression profiling on the basis of randomly assembled arrays of beads. Each bead carries a gene-specific probe sequence. There are multiple copies of each sequence-specific bead in an array, which contributes to measurement precision and reliability. We optimized the system for specific and sensitive analysis of mammalian RNA, and using RNA controls of defined concentration, obtained the following estimates of system performance: specificity of 1:250,000 in mammalian poly(A(+)) mRNA; limit of detection 0.13 pM; dynamic range 3.2 logs; and sufficient precision to detect 1.3-fold differences with 95% confidence within the dynamic range. Measurements of expression differences between human brain and liver were validated by concordance with quantitative real-time PCR (R(2) = 0.98 for log-transformed ratios, and slope of the best-fit line = 1.04, for 20 genes). Quantitative performance was further verified using a mouse B- and T-cell model system. We found published reports of B- or T-cell-specific expression for 42 of 59 genes that showed the greatest differential expression between B- and T-cells in our system. All of the literature observations were concordant with our results. Our experiments were carried out on a 96-array matrix system that requires only 100 ng of input RNA and uses standard microtiter plates to process samples in parallel. Our technology has advantages for analyzing multiple samples, is scalable to all known genes in a genome, and is flexible, allowing the use of standard or custom probes in an array.

- Quackenbush (2001) — computational analysis of microarray data.

  Computational analysis of microarray data. Quackenbush (2001). *Nature Reviews Genetics*. doi:10.1038/35076576

  > Microarray experiments are providing unprecedented quantities of genome-wide data on gene-expression patterns. Although this technique has been enthusiastically developed and applied in many biological contexts, the management and analysis of the millions of data points that result from these experiments has received less attention. Sophisticated computational tools are available, but the methods that are used to analyse the data can have a profound influence on the interpretation of the results. A basic understanding of these computational tools is therefore required for optimal experimental design and meaningful data analysis.

- Bolstad et al. (2003), Irizarry (2003) — quantile normalisation and RMA.

  A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bolstad et al. (2003). *Bioinformatics*. doi:10.1093/bioinformatics/19.2.185

Motivation When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations. Results We present three methods of performing normalization at the probe intensity level. These methods are called complete data methods because they make use of data from all arrays in an experiment to form the normalizing relation. These algorithms are compared to two methods that make use of a baseline array: a one number scaling based algorithm and a method that uses a non-linear normalizing relation by comparing the variability and bias of an expression measure. Two publicly available datasets are used to carry out the comparisons. The simplest and quickest complete data method is found to perform favorably. Availability Software implementing all three of the complete data normalization methods is available as part of the R package Affy, which is a part of the Bioconductor project http://www.bioconductor.org. Supplementary information Additional figures may be found at http://www.stat.berkeley.edu/~bolstad/normalize/index.html

Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Irizarry (2003). *Biostatistics*. doi:10.1093/biostatistics/4.2.249

In this paper we report exploratory analyses of high-density oligonucleotide array data from the Affymetrix GeneChip system with the objective of improving upon currently used measures of gene expression. Our analyses make use of three data sets: a small experimental study consisting of five MGU74A mouse GeneChip arrays, part of the data from an extensive spike-in study conducted by Gene Logic and Wyeth's Genetics Institute involving 95 HG-U95A human GeneChip arrays; and part of a dilution study conducted by Gene Logic involving 75 HG-U95A GeneChip arrays. We display some familiar features of the perfect match and mismatch probe (PM and MM) values of these data, and examine the variance-mean relationship with probe-level data from probes believed to be defective, and so delivering noise only. We explain why we need to normalize the arrays to one another using probe level intensities. We then examine the behavior of the PM and MM using spike-in data and assess three commonly used summary measures: Affymetrix's (i) average difference (AvDiff) and (ii) MAS 5.0 signal, and (iii) the Li and Wong multiplicative model-based expression index (MBEI). The exploratory data analyses of the probe level data motivate a new summary measure that is a robust multi-array average (RMA) of background-adjusted, normalized, and log-transformed PM values. We evaluate the four expression summary measures using the dilution study data, assessing their behavior in terms of bias, variance and (for MBEI and RMA) model fit. Finally, we evaluate the algorithms in terms of their ability to detect known levels of differential expression using the spike-in data. We conclude that there is no obvious downside to using RMA and attaching a standard error (SE) to this quantity using a linear model which removes probe-specific affinities.

- Smyth (2004) — limma: linear models and empirical Bayes.

Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Smyth (2004). *Statistical Applications in Genetics and Molecular Biology*. doi:10.2202/1544-6115.1027

The problem of identifying differentially expressed genes in designed microarray experiments is considered. Lonnstedt and Speed (2002) derived an expression for the posterior odds of differential expression in a replicated two-color experiment using a simple hierarchical parametric model. The purpose of this paper is to develop the hierarchical model of Lonnstedt and Speed (2002) into a practical approach for general microarray experiments with arbitrary numbers of treatments and RNA samples. The model is reset in the context of general linear models with arbitrary coefficients and contrasts of interest. The approach applies equally well to both single channel and two color microarray experiments. Consistent, closed form estimators are derived for the hyperparameters in the model. The estimators proposed have robust behavior even for small numbers of arrays and allow for incomplete data arising from spot filtering or spot quality weights. The posterior odds statistic is reformulated in terms of a moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations. The empirical Bayes approach is equivalent to shrinkage of the estimated sample variances towards a pooled estimate, resulting in far more stable inference when the number of arrays is small. The use of moderated t-statistics has the advantage over the posterior odds that the number of hyperparameters which need to estimated is reduced; in particular, knowledge of the non-null prior for the fold changes are not required. The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests of composite null hypotheses through the use of moderated F-statistics. The performance of the methods is demonstrated in a simulation study. Results are presented for two publicly available data sets.

- Eisen et al. (1998) — cluster analysis and display of genome-wide expression.

  Cluster analysis and display of genome-wide expression patterns. Eisen et al. (1998). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.95.25.14863

  > A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, conveying the clustering and the underlying expression data simultaneously in a form intuitive for biologists. We have found in the budding yeast Saccharomyces cerevisiae that clustering gene expression data groups together efficiently genes of known similar function, and we find a similar tendency in human data. Thus patterns seen in genome-wide expression experiments can be interpreted as indications of the status of cellular processes. Also, coexpression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of many genes for which information is not available currently.

- Golub et al. (1999), Perou et al. (2000), Veer et al. (2002) — classifying cancers by expression.

  Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Golub et al. (1999). *Science*. doi:10.1126/science.286.5439.531

  > Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

  Molecular portraits of human breast tumours. Perou et al. (2000). *Nature*. doi:10.1038/35021093

  > Human breast tumours are diverse in their natural history and in their responsiveness to treatments. Variation in transcriptional programs accounts for much of the biological diversity of human cells and tumours. In each cell, signal transduction and regulatory systems transduce information from the cell's identity to its environmental status, thereby controlling the level of expression of every gene in the genome. Here we have characterized variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals, using complementary DNA microarrays representing 8,102 human genes. These patterns provided a distinctive molecular portrait of each tumour. Twenty of the tumours were sampled twice, before and after a 16-week course of doxorubicin chemotherapy, and two tumours were paired with a lymph node metastasis from the same patient. Gene expression patterns in two tumour samples from the same individual were almost always more similar to each other than either was to any other sample. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by pervasive differences in their gene expression patterns.

  Gene expression profiling predicts clinical outcome of breast cancer. van 't Veer et al. (2002). *Nature*. doi:10.1038/415530a

  > Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70-80% of patients receiving this treatment would have survived without it. None of the signatures of breast cancer gene expression reported to date allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumour cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumours of BRCA1 carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and angiogenesis. This gene expression profile will outperform all currently used clinical parameters in predicting disease outcome. Our findings provide a strategy to select patients who would benefit from adjuvant therapy.

- Su et al. (2004) — a gene atlas of mouse and human tissues.

  A gene atlas of the mouse and human protein-encoding transcriptomes. Su et al. (2004). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0400782101

  > The tissue-specific pattern of mRNA expression can indicate important clues about gene function. High-density oligonucleotide arrays offer the opportunity to examine patterns of gene expression on a genome scale. Toward this end, we have designed custom arrays that interrogate the expression of the vast majority of protein-encoding human and mouse genes and have used them to profile a panel of 79 human and 61 mouse tissues. The resulting data set provides the expression patterns for thousands of predicted genes, as well as known and poorly characterized genes, from mice and humans. We have explored this data set for global trends in gene expression, evaluated commonly used lines of evidence in gene prediction methodologies, and investigated patterns indicative of chromosomal organization of transcription. We describe hundreds of regions of correlated transcription and show that some are subject to both tissue and parental allele-specific expression, suggesting a link between spatial expression and imprinting.

- Brazma et al. (2001), MAQC Consortium (2006) — standards and reproducibility.

  Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Brazma et al. (2001). *Nature Genetics*. doi:10.1038/ng1201-365

  > Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the Minimum Information About a Microarray Experiment (MIAME), that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

  The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. MAQC Consortium (2006). *Nature Biotechnology*. doi:10.1038/nbt1239

  > Over the last decade, the introduction of microarray technology has had a profound impact on gene expression research. The publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms to analyze identical RNA samples, has raised concerns about the reliability of this technology. The MicroArray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues. Expression data on four titration pools from two distinct reference RNA samples were generated at multiple test sites using a variety of microarray-based and alternative technology platforms. Here we describe the experimental design and probe mapping efforts behind the MAQC project. We show intraplatform consistency across test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed. This study provides a resource that represents an important first step toward establishing a framework for the use of microarrays in clinical and regulatory settings.

B.7.2 RNA-seq

- Mortazavi et al. (2008), Nagalakshmi et al. (2008) — the first RNA-seq studies.

  Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi et al. (2008). *Nature Methods*. doi:10.1038/nmeth.1226

  > We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41-52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude.

> Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3' untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are not readily measured by standard gene expression microarray or serial analysis of gene expression methods, were detected directly by mapping splice-crossing sequence reads. We observed 1.45 x 10(5) distinct splices, and alternative splices were prominent, with 3,500 different genes expressing one or more alternate internal splices.

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Nagalakshmi et al. (2008). *Science*. doi:10.1126/science.1158441

> The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3'-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

- Marioni et al. (2008) — RNA-seq compared with arrays.

  RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Marioni et al. (2008). *Genome Research*. doi:10.1101/gr.079558.108

  > Ultra-high-throughput sequencing is emerging as an attractive alternative to microarrays for genotyping, analysis of methylation patterns, and identification of transcription factor binding sites. Here, we describe an application of the Illumina sequencing (formerly Solexa sequencing) platform to study mRNA expression levels. Our goals were to estimate technical variance associated with Illumina sequencing in this context and to compare its ability to identify differentially expressed genes with existing array technologies. To do so, we estimated gene expression differences between liver and kidney RNA samples using multiple sequencing replicates, and compared the sequencing data to results obtained from Affymetrix arrays using the same RNA samples. We find that the Illumina sequencing data are highly replicable, with relatively little technical variation, and thus, for many purposes, it may suffice to sequence each mRNA sample only once (i.e., using one lane). The information in a single lane of Illumina sequencing data appears comparable to that in a single array in enabling identification of differentially expressed genes, while allowing for additional analyses such as detection of low-expressed genes, alternative splice variants, and novel transcripts. Based on our observations, we propose an empirical protocol and a statistical framework for the analysis of gene expression using ultra-high-throughput sequencing technology.

- Stark et al. (2019) — RNA sequencing: the teenage years.

  RNA sequencing: the teenage years. Stark et al. (2019). *Nature Reviews Genetics*. doi:10.1038/s41576-019-0150-2

  > Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

- Trapnell et al. (2009), Dobin et al. (2012), Kim et al. (2019) — spliced aligners (TopHat, STAR, HISAT2).

  TopHat: discovering splice junctions with RNA-Seq. Trapnell et al. (2009). *Bioinformatics*. doi:10.1093/bioinformatics/btp120

> Motivation A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, generates millions of short sequence fragments in a single run. These fragments, or 'reads', can be used to measure levels of gene expression and to identify novel splice variants of genes. However, current software for aligning RNA-Seq data to a genome relies on known splice junctions and cannot identify novel ones. TopHat is an efficient read-mapping algorithm designed to align reads from an RNA-Seq experiment to a reference genome without relying on known splice sites. Results We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20,000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. We describe several challenges unique to ab initio splice site discovery from RNA-Seq reads that will require further algorithm development. Availability TopHat is free, open-source software available from http://tophat.cbcb.umd.edu. Supplementary information Supplementary data are available at Bioinformatics online.

STAR: ultrafast universal RNA-seq aligner. Dobin et al. (2012). *Bioinformatics*. doi:10.1093/bioinformatics/bts635

> Motivation Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem because of the non-contiguous transcript structure, relatively short read lengths and constantly increasing throughput of the sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases. Results To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, aligning to the human genome 550 million 2 × 76 bp paired-end reads per hour on a modest 12-core server, while at the same time improving alignment sensitivity and precision. In addition to unbiased de novo detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of mapping full-length RNA sequences. Using Roche 454 sequencing of reverse transcription polymerase chain reaction amplicons, we experimentally validated 1960 novel intergenic splice junctions with an 80-90% success rate, corroborating the high precision of the STAR mapping strategy. Availability and implementation STAR is implemented as a standalone C++ code. STAR is free open source software distributed under GPLv3 license and can be downloaded from http://code.google.com/p/rna-star/.

Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Kim et al. (2019). *Nature Biotechnology*. doi:10.1038/s41587-019-0201-4

> The human reference genome represents only a small number of individuals, which limits its usefulness for genotyping. We present a method named HISAT2 (hierarchical indexing for spliced alignment of transcripts 2) that can align both DNA and RNA sequences using a graph Ferragina Manzini index. We use HISAT2 to represent and search an expanded model of the human reference genome in which over 14.5 million genomic variants in combination with haplotypes are incorporated into the data structure used for searching and alignment. We benchmark HISAT2 using simulated and real datasets to demonstrate that our strategy of representing a population of genomes, together with a fast, memory-efficient search algorithm, provides more detailed and accurate variant analyses than other methods. We apply HISAT2 for HLA typing and DNA fingerprinting; both applications form part of the HISAT-genotype software that enables analysis of haplotype-resolved genes or genomic regions. HISAT-genotype outperforms other computational methods and matches or exceeds the performance of laboratory-based assays.

- Trapnell et al. (2010), Pertea et al. (2015), Grabherr et al. (2011) — transcript assembly (Cufflinks, StringTie, Trinity).

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Trapnell et al. (2010). *Nature Biotechnology*. doi:10.1038/nbt.1621

> High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430

million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Pertea et al. (2015). *Nature Biotechnology*. doi:10.1038/nbt.3122

> Methods used to sequence the transcriptome often produce more than 200 million short sequences. We introduce StringTie, a computational method that applies a network flow algorithm originally developed in optimization theory, together with optional de novo assembly, to assemble these complex data sets into transcripts. When used to analyze both simulated and real data sets, StringTie produces more complete and accurate reconstructions of genes and better estimates of expression levels, compared with other leading transcript assembly programs including Cufflinks, IsoLasso, Scripture and Traph. For example, on 90 million reads from human blood, StringTie correctly assembled 10,990 transcripts, whereas the next best assembly was of 7,187 transcripts by Cufflinks, which is a 53% increase in transcripts assembled. On a simulated data set, StringTie correctly assembled 7,559 transcripts, which is 20% more than the 6,310 assembled by Cufflinks. As well as producing a more complete transcriptome assembly, StringTie runs faster on all data sets tested to date compared with other assembly software, including Cufflinks.

Full-length transcriptome assembly from RNA-Seq data without a reference genome. Grabherr et al. (2011). *Nature Biotechnology*. doi:10.1038/nbt.1883

> Massively parallel sequencing of cDNA has enabled deep and efficient probing of transcriptomes. Current approaches for transcript reconstruction from such data often rely on aligning reads to a reference genome, and are thus unsuitable for samples with a partial or missing reference genome. Here we present the Trinity method for de novo assembly of full-length transcripts and evaluate it on samples from fission yeast, mouse and whitefly, whose reference genome is not yet available. By efficiently constructing and analyzing sets of de Bruijn graphs, Trinity fully reconstructs a large fraction of transcripts, including alternatively spliced isoforms and transcripts from recently duplicated genes. Compared with other de novo transcriptome assemblers, Trinity recovers more full-length transcripts across a broad range of expression levels, with a sensitivity similar to methods that rely on genome alignments. Our approach provides a unified solution for transcriptome reconstruction in any sample, especially in the absence of a reference genome.

- Li and Dewey (2011) — RSEM and the EM algorithm.

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. Li & Dewey (2011). *BMC Bioinformatics*. doi:10.1186/1471-2105-12-323

> Background RNA-Seq is revolutionizing the way transcript abundances are measured. A key challenge in transcript quantification from RNA-Seq data is the handling of reads that map to multiple genes or isoforms. This issue is particularly important for quantification with de novo transcriptome assemblies in the absence of sequenced genomes, as it is difficult to determine which transcripts are isoforms of the same gene. A second significant issue is the design of RNA-Seq experiments, in terms of the number of reads, read length, and whether reads come from one or both ends of cDNA fragments. Results We present RSEM, an user-friendly software package for quantifying gene and isoform abundances from single-end or paired-end RNA-Seq data. RSEM outputs abundance estimates, 95% credibility intervals, and visualization files and can also simulate RNA-Seq data. In contrast to other existing tools, the software does not require a reference genome. Thus, in combination with a de novo transcriptome assembler, RSEM enables accurate transcript quantification for species without sequenced genomes. On simulated and real data sets, RSEM has superior or comparable performance to quantification methods that rely on a reference genome. Taking advantage of RSEM's ability to effectively use ambiguously-mapping reads, we show that accurate gene-level abundance estimates are best obtained with large numbers of short single-end reads. On the other hand, estimates of the relative frequencies of isoforms within single genes may be improved through the use of paired-end reads, depending on the number of possible splice forms for each gene. Conclusions RSEM is an accurate and user-friendly software tool for quantifying transcript abundances from RNA-Seq data. As it does not rely on the existence of a reference genome, it is particularly useful for quantification with de novo transcriptome assemblies. In addition, RSEM has

enabled valuable guidance for cost-efficient design of quantification experiments with RNA-Seq, which is currently relatively expensive.

- N. L. Bray et al. (2016), Patro et al. (2017) — kallisto and Salmon.

  Near-optimal probabilistic RNA-seq quantification. Bray et al. (2016). *Nature Biotechnology*. doi:10.1038/nbt.3519

  > We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. We use kallisto to analyze 30 million unaligned paired-end RNA-seq reads in <10 min on a standard laptop computer. This removes a major computational bottleneck in RNA-seq analysis.

  Salmon provides fast and bias-aware quantification of transcript expression. Patro et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4197

  > We introduce Salmon, a lightweight method for quantifying transcript abundance from RNA-seq reads. Salmon combines a new dual-phase parallel inference algorithm and feature-rich bias models with an ultra-fast read mapping procedure. It is the first transcriptome-wide quantifier to correct for fragment GC-content bias, which, as we demonstrate here, substantially improves the accuracy of abundance estimates and the sensitivity of subsequent differential expression analysis.

- Wagner et al. (2012) — why TPM rather than RPKM.

  Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Wagner et al. (2012). *Theory in Biosciences*. doi:10.1007/s12064-012-0162-3

  > Measures of RNA abundance are important for many areas of biology and often obtained from high-throughput RNA sequencing methods such as Illumina sequence data. These measures need to be normalized to remove technical biases inherent in the sequencing approach, most notably the length of the RNA species and the sequencing depth of a sample. These biases are corrected in the widely used reads per kilobase per million reads (RPKM) measure. Here, we argue that the intended meaning of RPKM is a measure of relative molar RNA concentration (rmc) and show that for each set of transcripts the average rmc is a constant, namely the inverse of the number of transcripts mapped. Further, we show that RPKM does not respect this invariance property and thus cannot be an accurate measure of rmc. We propose a slight modification of RPKM that eliminates this inconsistency and call it TPM for transcripts per million. TPM respects the average invariance and eliminates statistical biases inherent in the RPKM measure.

- Schurch et al. (2016) — how many replicates are needed.

  How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. Schurch et al. (2016). *RNA*. doi:10.1261/rna.053959.115

  > RNA-seq is now the technology of choice for genome-wide differential gene expression experiments, but it is not clear how many biological replicates are needed to ensure valid biological interpretation of the results or which statistical tools are best for analyzing the data. An RNA-seq experiment with 48 biological replicates in each of two conditions was performed to answer these questions and provide guidelines for experimental design. With three biological replicates, nine of the 11 tools evaluated found only 20%-40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates. This rises to >85% for the subset of SDE genes changing in expression by more than fourfold. To achieve >85% for all SDE genes regardless of fold change requires more than 20 biological replicates. The same nine tools successfully control their false discovery rate at ⊠5% for all numbers of replicates, while the remaining two tools fail to control their FDR adequately, particularly for low numbers of replicates. For future RNA-seq experiments, these results suggest that at least six biological replicates should be used, rising to at least 12 when it is important to identify SDE genes for all fold changes. If fewer than 12 replicates are used, a superior combination of true positive and false positive performances makes edgeR and DESeq2 the leading tools. For higher replicate numbers, minimizing false positives is more important and DESeq marginally outperforms the other tools.

- Leek et al. (2010) — batch effects.

  Tackling the widespread and critical impact of batch effects in high-throughput data. Leek et al. (2010). *Nature Reviews Genetics*. doi:10.1038/nrg2825

  > High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address. We review experimental and computational approaches for doing so.

B.7.3 Statistics and interpretation

- Oshlack et al. (2010) — from RNA-seq reads to differential expression.

  From RNA-seq reads to differential expression results. Oshlack et al. (2010). *Genome Biology*. doi:10.1186/gb-2010-11-12-220

  > Many methods and tools are available for preprocessing high-throughput RNA sequencing data and detecting differential expression.

- Robinson et al. (2009), Anders and Huber (2010) — edgeR and DESeq.

  edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Robinson et al. (2009). *Bioinformatics*. doi:10.1093/bioinformatics/btp616

  > It is expected that emerging digital gene expression (DGE) technologies will overtake microarray technologies in the near future for many functional genomics applications. One of the fundamental data analysis tasks, especially for gene expression studies, involves determining whether there is evidence that counts for a transcript or exon are significantly different across experimental conditions. edgeR is a Bioconductor software package for examining differential expression of replicated count data. An overdispersed Poisson model is used to account for both biological and technical variability. Empirical Bayes methods are used to moderate the degree of overdispersion across transcripts, improving the reliability of inference. The methodology can be used even with the most minimal levels of replication, provided at least one phenotype or experimental condition is replicated. The software may have other applications beyond sequencing data, such as proteome peptide count data. Availability The package is freely available under the LGPL licence from the Bioconductor web site (http://bioconductor.org).

  Differential expression analysis for sequence count data. Anders & Huber (2010). *Genome Biology*. doi:10.1186/gb-2010-11-10-r106

  > High-throughput sequencing assays such as RNA-Seq, ChIP-Seq or barcode counting provide quantitative readouts in the form of count data. To infer differential signal in such data correctly and with good statistical power, estimation of data variability throughout the dynamic range and a suitable error model are required. We propose a method based on the negative binomial distribution, with variance and mean linked by local regression and present an implementation, DESeq, as an R/Bioconductor package.

- Law et al. (2014) — voom.

  voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Law et al. (2014). *Genome Biology*. doi:10.1186/gb-2014-15-2-r29

  > New normal linear modeling strategies are presented for analyzing read counts from RNA-seq experiments. The voom method estimates the mean-variance relationship of the log-counts, generates a precision weight for each observation and enters these into the limma empirical Bayes analysis pipeline. This opens access for RNA-seq analysts to a large body of methodology developed for microarrays. Simulation studies show that voom performs as well or better than count-based RNA-seq methods even when the data are generated according to the assumptions of the earlier methods. Two case studies illustrate the use of linear modeling and gene set testing methods.

- Storey and Tibshirani (2003), Efron and Tibshirani (2002) — q-values and local FDR.

  Statistical significance for genomewide studies. Storey & Tibshirani (2003). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1530509100

  > With the increase in genomewide experiments and the sequencing of multiple genomes, the analysis of large data sets has become commonplace in biology. It is often the case that thousands of features in a genomewide data set are tested against some null hypothesis, where a number of features are expected to be significant. Here we propose an approach to measuring statistical significance in these genomewide studies based on the concept of the false discovery rate. This approach offers a sensible balance between the number of true and false positives that is automatically calibrated and easily interpreted. In doing so, a measure of statistical significance called the q value is associated with each tested feature. The q value is similar to the well known p value, except it is a measure of significance in terms of the false discovery rate rather than the false positive rate. Our approach avoids a flood of false positive results, while offering a more liberal criterion than what has been used in genome scans for linkage.

  Empirical bayes methods and false discovery rates for microarrays. Efron & Tibshirani (2002). *Genetic Epidemiology*. doi:10.1002/gepi.1124

  > In a classic two-sample problem, one might use Wilcoxon's statistic to test for a difference between treatment and control subjects. The analogous microarray experiment yields thousands of Wilcoxon statistics, one for each gene on the array, and confronts the statistician with a difficult simultaneous inference situation. We will discuss two inferential approaches to this problem: an empirical Bayes method that requires very little a priori Bayesian modeling, and the frequentist method of "false discovery rates" proposed by Benjamini and Hochberg in 1995. It turns out that the two methods are closely related and can be used together to produce sensible simultaneous inferences.

- Krzywinski and Altman (2013), Nuzzo (2014), Wasserstein and Lazar (2016) — the importance of being uncertain; p-values and their misuse.

  Significance, P values and t-tests. Krzywinski & Altman (2013). *Nature Methods*. doi:10.1038/nmeth.2698

  Scientific method: Statistical errors. Nuzzo (2014). *Nature*. doi:10.1038/506150a

  The ASA Statement on p -Values: Context, Process, and Purpose. Wasserstein & Lazar (2016). *The American Statistician*. doi:10.1080/00031305.2016.1154108

- Noble (2009) — how does multiple testing correction work?

  How does multiple testing correction work?. Noble (2009). *Nature Biotechnology*. doi:10.1038/nbt1209-1135

- Holm (1979) — Holm's step-down procedure.

  A simple sequentially rejective multiple test procedure. Holm (1979). *Scandinavian Journal of Statistics*.

- Ioannidis (2005) — why most published research findings are false.

  Why Most Published Research Findings Are False. Ioannidis (2005). *PLoS Medicine*. doi:10.1371/journal.pmed.0020124

  > There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current

> scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

- Lever et al. (2017) — principal component analysis.

  Principal component analysis. Lever et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4346

- Maaten and Hinton (2008), McInnes et al. (2018), Becht et al. (2018) — t-SNE and UMAP.

  Visualizing Data using t-SNE. van der Maaten & Hinton (2008). *Journal of Machine Learning Research*. https://www.jmlr.org/papers/v9/vandermaaten08a.html

  UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. McInnes et al. (2018). https://arxiv.org/abs/1802.03426

  Dimensionality reduction for visualizing single-cell data using UMAP. Becht et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4314

  > Advances in single-cell technologies have enabled high-resolution dissection of tissue composition. Several tools for dimensionality reduction are available to analyze the large number of parameters generated in single-cell studies. Recently, a nonlinear dimensionality-reduction technique, uniform manifold approximation and projection (UMAP), was developed for the analysis of any type of high-dimensional data. Here we apply it to biological data, using three well-characterized mass cytometry and single-cell RNA sequencing datasets. Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters. The work highlights the use of UMAP for improved visualization and interpretation of single-cell data.

- Chari and Pachter (2023) — the specious art of single-cell genomics.

  The specious art of single-cell genomics. Chari & Pachter (2023). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1011288

  > Dimensionality reduction is standard practice for filtering noise and identifying relevant features in large-scale data analyses. In biology, single-cell genomics studies typically begin with reduction to 2 or 3 dimensions to produce "all-in-one" visuals of the data that are amenable to the human eye, and these are subsequently used for qualitative and quantitative exploratory analysis. However, there is little theoretical support for this practice, and we show that extreme dimension reduction, from hundreds or thousands of dimensions to 2, inevitably induces significant distortion of high-dimensional datasets. We therefore examine the practical implications of low-dimensional embedding of single-cell data and find that extensive distortions and inconsistent practices make such embeddings counter-productive for exploratory, biological analyses. In lieu of this, we discuss alternative approaches for conducting targeted embedding and feature exploration to enable hypothesis-driven biological discovery.

- Subramanian et al. (2005) — gene set enrichment analysis.

  Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Subramanian et al. (2005). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0506580102

  > Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

- Huang et al. (2008), Liberzon et al. (2015) — DAVID and MSigDB.

  Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Huang et al. (2008). *Nature Protocols*. doi:10.1038/nprot.2008.211

  > DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high-throughput and integrated data-mining environment, to analyze gene lists derived from high-throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway-mining tools such as gene functional classification, functional annotation chart or clustering and functional annotation table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies.

  The Molecular Signatures Database Hallmark Gene Set Collection. Liberzon et al. (2015). *Cell Systems*. doi:10.1016/j.cels.2015.12.004

  > The Molecular Signatures Database (MSigDB) is one of the most widely used and comprehensive databases of gene sets for performing gene set enrichment analysis. Since its creation, MSigDB has grown beyond its roots in metabolic disease and cancer to include >10,000 gene sets. These better represent a wider range of biological processes and diseases, but the utility of the database is reduced by increased redundancy across, and heterogeneity within, gene sets. To address this challenge, here we use a combination of automated approaches and expert curation to develop a collection of ”hallmark” gene sets as part of MSigDB. Each hallmark in this collection consists of a ”refined” gene set, derived from multiple ”founder” sets, that conveys a specific biological state or process and displays coherent expression. The hallmarks effectively summarize most of the relevant information of the original founder sets and, by reducing both variation and redundancy, provide more refined and concise inputs for gene set enrichment analysis.

- Langfelder and Horvath (2008) — WGCNA co-expression networks.

  WGCNA: an R package for weighted correlation network analysis. Langfelder & Horvath (2008). *BMC Bioinformatics*. doi:10.1186/1471-2105-9-559

  > Background Correlation networks are increasingly being used in bioinformatics applications. For example, weighted gene co-expression network analysis is a systems biology method for describing the correlation patterns among genes across microarray samples. Weighted correlation network analysis (WGCNA) can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module eigengene or an intramodular hub gene, for relating modules to one another and to external sample traits (using eigengene network methodology), and for calculating module membership measures. Correlation networks facilitate network based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, e.g. cancer, mouse genetics, yeast genetics, and analysis of brain imaging data. While parts of the correlation network methodology have been described in separate publications, there is a need to provide a user-friendly, comprehensive, and consistent software implementation and an accompanying tutorial. Results The WGCNA R software package is a comprehensive collection of R functions for performing various aspects of weighted correlation network analysis. The package includes functions for network construction, module detection, gene selection, calculations of topological properties, data simulation, visualization, and interfacing with external software. Along with the R package we also present R software tutorials. While the methods development was motivated by gene expression data, the underlying data mining approach can be applied to a variety of different settings. Conclusion The WGCNA package provides R functions for weighted correlation network analysis, e.g. co-expression network analysis of gene expression data. The R package along with its source code and additional material are freely available at http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA.

- The GTEx Consortium (2020) — GTEx: genetic effects on expression across tissues.

  The GTEx Consortium atlas of genetic regulatory effects across human tissues. The GTEx Consortium (2020). *Science*. doi:10.1126/science.aaz1776

  > The Genotype-Tissue Expression (GTEx) project was established to characterize genetic effects on the transcriptome across human tissues and to link these regulatory mechanisms to trait and disease associations. Here, we present analyses of the version 8 data, examining 15,201 RNA-sequencing samples from 49 tissues of 838 postmortem donors. We comprehensively characterize genetic associations for gene

> expression and splicing in cis and trans, showing that regulatory associations are found for almost all genes, and describe the underlying molecular mechanisms and their contribution to allelic heterogeneity and pleiotropy of complex traits. Leveraging the large diversity of tissues, we provide insights into the tissue specificity of genetic effects and show that cell type composition is a key factor in understanding gene regulatory mechanisms in human tissues.

- Schwanhäusser et al. (2011), Liu et al. (2016) — how mRNA levels relate to protein levels.

  Global quantification of mammalian gene expression control. Schwanhäusser et al. (2011). *Nature*. doi:10.1038/nature10098

  > Gene expression is a multistep process that involves the transcription, translation and turnover of messenger RNAs and proteins. Although it is one of the most fundamental processes of life, the entire cascade has never been quantified on a genome-wide scale. Here we simultaneously measured absolute mRNA and protein abundance and turnover by parallel metabolic pulse labelling for more than 5,000 genes in mammalian cells. Whereas mRNA and protein levels correlated better than previously thought, corresponding half-lives showed no correlation. Using a quantitative model we have obtained the first genome-scale prediction of synthesis rates of mRNAs and proteins. We find that the cellular abundance of proteins is predominantly controlled at the level of translation. Genes with similar combinations of mRNA and protein stability shared functional properties, indicating that half-lives evolved under energetic and dynamic constraints. Quantitative information about all stages of gene expression provides a rich resource and helps to provide a greater understanding of the underlying design principles.

  On the Dependency of Cellular Protein Levels on mRNA Abundance. Liu et al. (2016). *Cell*. doi:10.1016/j.cell.2016.03.014

  > The question of how genomic information is expressed to determine phenotypes is of central importance for basic and translational life science research and has been studied by transcriptomic and proteomic profiling. Here, we review the relationship between protein and mRNA levels under various scenarios, such as steady state, long-term state changes, and short-term adaptation, demonstrating the complexity of gene expression regulation, especially during dynamic transitions. The spatial and temporal variations of mRNAs, as well as the local availability of resources for protein biosynthesis, strongly influence the relationship between protein levels and their coding transcripts. We further discuss the buffering of mRNA fluctuations at the level of protein concentrations. We conclude that transcript levels by themselves are not sufficient to predict protein levels in many scenarios and to thus explain genotype-phenotype relationships and that high-quality data quantifying different levels of gene expression are indispensable for the complete understanding of biological processes.

- Uhlén et al. (2015) — the Human Protein Atlas.

  Tissue-based map of the human proteome. Uhlén et al. (2015). *Science*. doi:10.1126/science.1260419

  > Resolving the molecular details of proteome variation in the different tissues and organs of the human body will greatly increase our knowledge of human biology and disease. Here, we present a map of the human tissue proteome based on an integrated omics approach that involves quantitative transcriptomics at the tissue and organ level, combined with tissue microarray-based immunohistochemistry, to achieve spatial localization of proteins down to the single-cell level. Our tissue-based analysis detected more than 90% of the putative protein-coding genes. We used this approach to explore the human secretome, the membrane proteome, the druggable proteome, the cancer proteome, and the metabolic functions in 32 different tissues and organs. All the data are integrated in an interactive Web-based database that allows exploration of individual proteins, as well as navigation of global expression patterns, in all major tissues and organs in the human body.

B.7.4 Single-cell and spatial genomics

- Heumos et al. (2023) — best practices for single-cell analysis across modalities.

  Best practices for single-cell analysis across modalities. Heumos et al. (2023). *Nature Reviews Genetics*. doi:10.1038/s41576-023-00586-w

> Recent advances in single-cell technologies have enabled high-throughput molecular profiling of cells across modalities and locations. Single-cell transcriptomics data can now be complemented by chromatin accessibility, surface protein expression, adaptive immune receptor repertoire profiling and spatial information. The increasing availability of single-cell data across modalities has motivated the development of novel computational methods to help analysts derive biological insights. As the field grows, it becomes increasingly difficult to navigate the vast landscape of tools and analysis steps. Here, we summarize independent benchmarking studies of unimodal and multimodal single-cell analysis across modalities to suggest comprehensive best-practice workflows for the most common analysis steps. Where independent benchmarks are not available, we review and contrast popular methods. Our article serves as an entry point for novices in the field of single-cell (multi-)omic analysis and guides advanced users to the most recent best practices.

- Tang et al. (2009) — the first single-cell transcriptome.

  mRNA-Seq whole-transcriptome analysis of a single cell. Tang et al. (2009). *Nature Methods*. doi:10.1038/nmeth.1315

  > Next-generation sequencing technology is a powerful tool for transcriptome analysis. However, under certain conditions, only a small amount of material is available, which requires more sensitive techniques that can preferably be used at the single-cell level. Here we describe a single-cell digital gene expression profiling assay. Using our mRNA-Seq assay with only a single mouse blastomere, we detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads. Moreover, 8-19% of the genes with multiple known transcript isoforms expressed at least two isoforms in the same blastomere or oocyte, which unambiguously demonstrated the complexity of the transcript variants at whole-genome scale in individual cells. Finally, for Dicer1(-/-) and Ago2(-/-) (Eif2c2(-/-)) oocytes, we found that 1,696 and 1,553 genes, respectively, were abnormally upregulated compared to wild-type controls, with 619 genes in common.

- Svensson et al. (2018) — exponential scaling of single-cell RNA-seq.

  Exponential scaling of single-cell RNA-seq in the past decade. Svensson et al. (2018). *Nature Protocols*. doi:10.1038/nprot.2017.149

  > Measurement of the transcriptomes of single cells has been feasible for only a few years, but it has become an extremely popular assay. While many types of analysis can be carried out and various questions can be answered by single-cell RNA-seq, a central focus is the ability to survey the diversity of cell types in a sample. Unbiased and reproducible cataloging of gene expression patterns in distinct cell types requires large numbers of cells. Technological developments and protocol improvements have fueled consistent and exponential increases in the number of cells that can be studied in single-cell RNA-seq analyses. In this Perspective, we highlight the key technological developments that have enabled this growth in the data obtained from single-cell RNA-seq experiments.

- Picelli et al. (2013), Macosko et al. (2015), Klein et al. (2015), Zheng et al. (2017) — Smart-seq2 and droplet methods.

  Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Picelli et al. (2013). *Nature Methods*. doi:10.1038/nmeth.2639

  > Single-cell gene expression analyses hold promise for characterizing cellular heterogeneity, but current methods compromise on either the coverage, the sensitivity or the throughput. Here, we introduce Smart-seq2 with improved reverse transcription, template switching and preamplification to increase both yield and length of cDNA libraries generated from individual cells. Smart-seq2 transcriptome libraries have improved detection, coverage, bias and accuracy compared to Smart-seq libraries and are generated with off-the-shelf reagents at lower cost.

  Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Macosko et al. (2015). *Cell*. doi:10.1016/j.cell.2015.05.002

  > Cells, the basic units of biological structure and function, vary broadly in type and state. Single-cell genomics can characterize cell identity and function, but limitations of ease and scale have prevented its broad application. Here we describe Drop-seq, a strategy for quickly profiling thousands of individual cells by separating them into nanoliter-sized aqueous droplets, associating a different barcode with each cell's RNAs, and sequencing them all together. Drop-seq analyzes mRNA transcripts from thousands of

> individual cells simultaneously while remembering transcripts' cell of origin. We analyzed transcriptomes from 44,808 mouse retinal cells and identified 39 transcriptionally distinct cell populations, creating a molecular atlas of gene expression for known retinal cell classes and novel candidate cell subtypes. Drop-seq will accelerate biological discovery by enabling routine transcriptional profiling at single-cell resolution. VIDEO ABSTRACT.

Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Klein et al. (2015). *Cell*. doi:10.1016/j.cell.2015.04.044

> It has long been the dream of biologists to map gene expression at the single-cell level. With such data one might track heterogeneous cell sub-populations, and infer regulatory relationships between genes and pathways. Recently, RNA sequencing has achieved single-cell resolution. What is limiting is an effective way to routinely isolate and process large numbers of individual cells for quantitative in-depth sequencing. We have developed a high-throughput droplet-microfluidic approach for barcoding the RNA from thousands of individual cells for subsequent analysis by next-generation sequencing. The method shows a surprisingly low noise profile and is readily adaptable to other sequencing-based assays. We analyzed mouse embryonic stem cells, revealing in detail the population structure and the heterogeneous onset of differentiation after leukemia inhibitory factor (LIF) withdrawal. The reproducibility of these high-throughput single-cell data allowed us to deconstruct cell populations and infer gene expression relationships. VIDEO ABSTRACT.

Massively parallel digital transcriptional profiling of single cells. Zheng et al. (2017). *Nature Communications*. doi:10.1038/ncomms14049

> Characterizing the transcriptome of individual cells is fundamental to understanding complex biological systems. We describe a droplet-based system that enables 3' mRNA counting of tens of thousands of single cells per sample. Cell encapsulation, of up to 8 samples at a time, takes place in ~6 min, with ~50% cell capture efficiency. To demonstrate the system's technical performance, we collected transcriptome data from ~250k single cells across 29 samples. We validated the sensitivity of the system and its ability to detect rare populations using cell lines and synthetic RNAs. We profiled 68k peripheral blood mononuclear cells to demonstrate the system's ability to characterize large immune populations. Finally, we used sequence variation in the transcriptome data to determine host and donor chimerism at single-cell resolution from bone marrow mononuclear cells isolated from transplant patients.

- Islam et al. (2013) — unique molecular identifiers.

Quantitative single-cell RNA-seq with unique molecular identifiers. Islam et al. (2013). *Nature Methods*. doi:10.1038/nmeth.2772

> Single-cell RNA sequencing (RNA-seq) is a powerful tool to reveal cellular heterogeneity, discover new cell types and characterize tumor microevolution. However, losses in cDNA synthesis and bias in cDNA amplification lead to severe quantitative errors. We show that molecular labels--random sequences that label individual molecules--can nearly eliminate amplification noise, and that microfluidic sample preparation and optimized reagents produce a fivefold improvement in mRNA capture efficiency.

- Svensson (2020) — droplet scRNA-seq is not zero-inflated.

Droplet scRNA-seq is not zero-inflated. Svensson (2020). *Nature Biotechnology*. doi:10.1038/s41587-019-0379-5

- Lun et al. (2016), Hafemeister and Satija (2019) — normalisation of single-cell data.

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Lun et al. (2016). *Genome Biology*. doi:10.1186/s13059-016-0947-7

> Normalization of single-cell RNA sequencing data is necessary to eliminate cell-specific biases prior to downstream analyses. However, this is not straightforward for noisy single-cell data where many counts are zero. We present a novel approach where expression values are summed across pools of cells, and the summed values are used for normalization. Pool-based size factors are then deconvolved to yield cell-based factors. Our deconvolution approach outperforms existing methods for accurate normalization of cell-specific biases in simulated data. Similar behavior is observed in real data, where deconvolution improves the relevance of results of downstream analyses.

Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Hafemeister & Satija (2019). *Genome Biology*. doi:10.1186/s13059-019-1874-1

> Single-cell RNA-seq (scRNA-seq) data exhibits significant cell-to-cell variation due to technical factors, including the number of molecules detected in each cell, which can confound biological heterogeneity with technical effects. To address this, we present a modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiments. We propose that the Pearson residuals from "regularized negative binomial regression," where cellular sequencing depth is utilized as a covariate in a generalized linear model, successfully remove the influence of technical characteristics from downstream analyses while preserving biological heterogeneity. Importantly, we show that an unconstrained negative binomial model may overfit scRNA-seq data, and overcome this by pooling information across genes with similar abundances to obtain stable parameter estimates. Our procedure omits the need for heuristic steps including pseudocount addition or log-transformation and improves common downstream analytical tasks such as variable gene selection, dimensional reduction, and differential expression. Our approach can be applied to any UMI-based scRNA-seq dataset and is freely available as part of the R package sctransform, with a direct interface to our single-cell toolkit Seurat.

- Traag et al. (2019) — the Leiden algorithm.

From Louvain to Leiden: guaranteeing well-connected communities. Traag et al. (2019). *Scientific Reports*. doi:10.1038/s41598-019-41695-z

> Community detection is often used to understand the structure of large and complex networks. One of the most popular algorithms for uncovering community structure is the so-called Louvain algorithm. We show that this algorithm has a major defect that largely went unnoticed until now: the Louvain algorithm may yield arbitrarily badly connected communities. In the worst case, communities may even be disconnected, especially when running the algorithm iteratively. In our experimental analysis, we observe that up to 25% of the communities are badly connected and up to 16% are disconnected. To address this problem, we introduce the Leiden algorithm. We prove that the Leiden algorithm yields communities that are guaranteed to be connected. In addition, we prove that, when the Leiden algorithm is applied iteratively, it converges to a partition in which all subsets of all communities are locally optimally assigned. Furthermore, by relying on a fast local move approach, the Leiden algorithm runs faster than the Louvain algorithm. We demonstrate the performance of the Leiden algorithm for several benchmark and real-world networks. We find that the Leiden algorithm is faster than the Louvain algorithm and uncovers better partitions, in addition to providing explicit guarantees.

- Hao et al. (2021), Wolf et al. (2018) — Seurat and Scanpy.

Integrated analysis of multimodal single-cell data. Hao et al. (2021). *Cell*. doi:10.1016/j.cell.2021.04.048

> The simultaneous measurement of multiple modalities represents an exciting frontier for single-cell genomics and necessitates computational methods that can define cellular states based on multimodal data. Here, we introduce "weighted-nearest neighbor" analysis, an unsupervised framework to learn the relative utility of each data type in each cell, enabling an integrative analysis of multiple modalities. We apply our procedure to a CITE-seq dataset of 211,000 human peripheral blood mononuclear cells (PBMCs) with panels extending to 228 antibodies to construct a multimodal reference atlas of the circulating immune system. Multimodal analysis substantially improves our ability to resolve cell states, allowing us to identify and validate previously unreported lymphoid subpopulations. Moreover, we demonstrate how to leverage this reference to rapidly map new datasets and to interpret immune responses to vaccination and coronavirus disease 2019 (COVID-19). Our approach represents a broadly applicable strategy to analyze single-cell multimodal datasets and to look beyond the transcriptome toward a unified and multimodal definition of cellular identity.

SCANPY: large-scale single-cell gene expression data analysis. Wolf et al. (2018). *Genome Biology*. doi:10.1186/s13059-017-1382-0

> SCANPY is a scalable toolkit for analyzing single-cell gene expression data. It includes methods for preprocessing, visualization, clustering, pseudotime and trajectory inference, differential expression testing, and simulation of gene regulatory networks. Its Python-based implementation efficiently deals with data sets of more than one million cells (https://github.com/theislab/Scanpy). Along with SCANPY, we present ANNDATA, a generic class for handling annotated data matrices (https://github.com/theislab/anndata).

- Korsunsky et al. (2019), Lopez et al. (2018) — data integration (Harmony, scVI).

Fast, sensitive and accurate integration of single-cell data with Harmony. Korsunsky et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0619-0

> The emerging diversity of single-cell RNA-seq datasets allows for the full transcriptional characterization of cell types across a wide variety of biological and clinical conditions. However, it is challenging to analyze them together, particularly when datasets are assayed with different technologies, because biological and technical differences are interspersed. We present Harmony (https://github.com/immunogenomics/harmony), an algorithm that projects cells into a shared embedding in which cells group by cell type rather than dataset-specific conditions. Harmony simultaneously accounts for multiple experimental and biological factors. In six analyses, we demonstrate the superior performance of Harmony to previously published algorithms while requiring fewer computational resources. Harmony enables the integration of ~10 6 cells on a personal computer. We apply Harmony to peripheral blood mononuclear cells from datasets with large experimental differences, five studies of pancreatic islet cells, mouse embryogenesis datasets and the integration of scRNA-seq with spatial transcriptomics data.

Deep generative modeling for single-cell transcriptomics. Lopez et al. (2018). *Nature Methods*. doi:10.1038/s41592-018-0229-2

> Single-cell transcriptome measurements can reveal unexplored biological diversity, but they suffer from technical noise and bias that must be modeled to account for the resulting uncertainty in downstream analyses. Here we introduce single-cell variational inference (scVI), a ready-to-use scalable framework for the probabilistic representation and analysis of gene expression in single cells (https://github.com/YosefLab/scVI). scVI uses stochastic optimization and deep neural networks to aggregate information across similar cells and genes and to approximate the distributions that underlie observed expression values, while accounting for batch effects and limited sensitivity. We used scVI for a range of fundamental analysis tasks including batch correction, visualization, clustering, and differential expression, and achieved high accuracy for each task.

- Trapnell et al. (2014), La Manno et al. (2018) — pseudotime and RNA velocity.

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Trapnell et al. (2014). *Nature Biotechnology*. doi:10.1038/nbt.2859

> Defining the transcriptional dynamics of a temporal process such as cell differentiation is challenging owing to the high variability in gene expression between individual cells. Time-series gene expression analyses of bulk cells have difficulty distinguishing early and late phases of a transcriptional cascade or identifying rare subpopulations of cells, and single-cell proteomic methods rely on a priori knowledge of key distinguishing markers. Here we describe Monocle, an unsupervised algorithm that increases the temporal resolution of transcriptome dynamics using single-cell RNA-Seq data collected at multiple time points. Applied to the differentiation of primary human myoblasts, Monocle revealed switch-like changes in expression of key regulatory factors, sequential waves of gene regulation, and expression of regulators that were not known to act in differentiation. We validated some of these predicted regulators in a loss-of function screen. Monocle can in principle be used to recover single-cell gene expression kinetics from a wide array of cellular processes, including differentiation, proliferation and oncogenic transformation.

RNA velocity of single cells. La Manno et al. (2018). *Nature*. doi:10.1038/s41586-018-0414-6

> RNA abundance is a powerful indicator of the state of individual cells. Single-cell RNA sequencing can reveal RNA abundance with high quantitative accuracy, sensitivity and throughput 1 . However, this approach captures only a static snapshot at a point in time, posing a challenge for the analysis of time-resolved phenomena such as embryogenesis or tissue regeneration. Here we show that RNA velocity-the time derivative of the gene expression state-can be directly estimated by distinguishing between unspliced and spliced mRNAs in common single-cell RNA sequencing protocols. RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours. We validate its accuracy in the neural crest lineage, demonstrate its use on multiple published datasets and technical platforms, reveal the branching lineage tree of the developing mouse hippocampus, and examine the kinetics of transcription in human embryonic brain. We expect RNA velocity to greatly aid the analysis of developmental lineages and cellular dynamics, particularly in humans.

- Ståhl et al. (2016), Rodriques et al. (2019), Chen et al. (2015), Marx (2021) — spatial transcriptomics.

  Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Ståhl et al. (2016). *Science*. doi:10.1126/science.aaf2403

  > Analysis of the pattern of proteins or messengerRNAs (mRNAs) in histological tissue sections is a cornerstone in biomedical research and diagnostics. This typically involves the visualization of a few proteins or expressed genes at a time. We have devised a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue sections. By positioning histological sections on arrayed reverse transcription primers with unique positional barcodes, we demonstrate high-quality RNA-sequencing data with maintained two-dimensional positional information from the mouse brain and human breast cancer. Spatial transcriptomics provides quantitative gene expression data and visualization of the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics.

  Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Rodriques et al. (2019). *Science*. doi:10.1126/science.aaw1219

  > Spatial positions of cells in tissues strongly influence function, yet a high-throughput, genome-wide readout of gene expression with cellular resolution is lacking. We developed Slide-seq, a method for transferring RNA from tissue sections onto a surface covered in DNA-barcoded beads with known positions, allowing the locations of the RNA to be inferred by sequencing. Using Slide-seq, we localized cell types identified by single-cell RNA sequencing datasets within the cerebellum and hippocampus, characterized spatial gene expression patterns in the Purkinje layer of mouse cerebellum, and defined the temporal evolution of cell type-specific responses in a mouse model of traumatic brain injury. These studies highlight how Slide-seq provides a scalable method for obtaining spatially resolved gene expression data at resolutions comparable to the sizes of individual cells.

  Spatially resolved, highly multiplexed RNA profiling in single cells. Chen et al. (2015). *Science*. doi:10.1126/science.aaa6090

  > Knowledge of the expression profile and spatial landscape of the transcriptome in individual cells is essential for understanding the rich repertoire of cellular behaviors. Here, we report multiplexed error-robust fluorescence in situ hybridization (MERFISH), a single-molecule imaging approach that allows the copy numbers and spatial localizations of thousands of RNA species to be determined in single cells. Using error-robust encoding schemes to combat single-molecule labeling and detection errors, we demonstrated the imaging of 100 to 1000 distinct RNA species in hundreds of individual cells. Correlation analysis of the ~10(4) to 10(6) pairs of genes allowed us to constrain gene regulatory networks, predict novel functions for many unannotated genes, and identify distinct spatial distribution patterns of RNAs that correlate with properties of the encoded proteins.

  Method of the Year: spatially resolved transcriptomics. Marx (2021). *Nature Methods*. doi:10.1038/s41592-020-01033-y

- Theodoris et al. (2023), Cui et al. (2024) — single-cell foundation models.

  Transfer learning enables predictions in network biology. Theodoris et al. (2023). *Nature*. doi:10.1038/s41586-023-06139-9

  > Mapping gene networks requires large amounts of transcriptomic data to learn the connections between genes, which impedes discoveries in settings with limited data, including rare diseases and diseases affecting clinically inaccessible tissues. Recently, transfer learning has revolutionized fields such as natural language understanding 1,2 and computer vision 3 by leveraging deep learning models pretrained on large-scale general datasets that can then be fine-tuned towards a vast array of downstream tasks with limited task-specific data. Here, we developed a context-aware, attention-based deep learning model, Geneformer, pretrained on a large-scale corpus of about 30 million single-cell transcriptomes to enable context-specific predictions in settings with limited data in network biology. During pretraining, Geneformer gained a fundamental understanding of network dynamics, encoding network hierarchy in the attention weights of the model in a completely self-supervised manner. Fine-tuning towards a diverse panel of downstream tasks relevant to chromatin and network dynamics using limited task-specific data demonstrated that Geneformer consistently boosted predictive accuracy. Applied to disease modelling with limited patient data, Geneformer identified candidate therapeutic targets for cardiomyopathy. Overall, Geneformer represents a pretrained deep learning model from which fine-tuning towards a broad range of

downstream applications can be pursued to accelerate discovery of key network regulators and candidate therapeutic targets.

scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Cui et al. (2024). *Nature Methods*. doi:10.1038/s41592-024-02201-0

> Generative pretrained models have achieved remarkable success in various domains such as language and computer vision. Specifically, the combination of large-scale diverse datasets and pretrained transformers has emerged as a promising approach for developing foundation models. Drawing parallels between language and cellular biology (in which texts comprise words; similarly, cells are defined by genes), our study probes the applicability of foundation models to advance cellular biology and genetic research. Using burgeoning single-cell sequencing data, we have constructed a foundation model for single-cell biology, scGPT, based on a generative pretrained transformer across a repository of over 33 million cells. Our findings illustrate that scGPT effectively distills critical biological insights concerning genes and cells. Through further adaptation of transfer learning, scGPT can be optimized to achieve superior performance across diverse downstream applications. This includes tasks such as cell type annotation, multi-batch integration, multi-omic integration, perturbation response prediction and gene network inference.

- Kedzierska et al. (2025), Ahlmann-Eltze et al. (2025) — evaluating foundation models against simple baselines.

Zero-shot evaluation reveals limitations of single-cell foundation models. Kedzierska et al. (2025). *Genome Biology*. doi:10.1186/s13059-025-03574-x

> Foundation models such as scGPT and Geneformer have not been rigorously evaluated in a setting where they are used without any further training (i.e., zero-shot). Understanding the performance of models in zero-shot settings is critical to applications that exclude the ability to fine-tune, such as discovery settings where labels are unknown. Our evaluation of the zero-shot performance of Geneformer and scGPT suggests that, in some cases, these models may face reliability challenges and could be outperformed by simpler methods. Our findings underscore the importance of zero-shot evaluations in development and deployment of foundation models in single-cell research.

Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Ahlmann-Eltze et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02772-6

> Recent research in deep-learning-based foundation models promises to learn representations of single-cell data that enable prediction of the effects of genetic perturbations. Here we compared five foundation models and two other deep learning models against deliberately simple baselines for predicting transcriptome changes after single or double perturbations. None outperformed the baselines, which highlights the importance of critical benchmarking in directing and evaluating method development.

- Regev et al. (2017), The Tabula Muris Consortium (2018), The Tabula Sapiens Consortium (2022) — cell atlases.

The Human Cell Atlas. Regev et al. (2017). *eLife*. doi:10.7554/elife.27041

> The recent advent of methods for high-throughput single-cell molecular profiling has catalyzed a growing sense in the scientific community that the time is ripe to complete the 150-year-old effort to identify all cell types in the human body. The Human Cell Atlas Project is an international collaborative effort that aims to define all human cell types in terms of distinctive molecular profiles (such as gene expression profiles) and to connect this information with classical cellular descriptions (such as location and morphology). An open comprehensive reference map of the molecular state of cells in healthy human tissues would propel the systematic study of physiological states, developmental trajectories, regulatory circuitry and interactions of cells, and also provide a framework for understanding cellular dysregulation in human disease. Here we describe the idea, its potential utility, early proofs-of-concept, and some design considerations for the Human Cell Atlas, including a commitment to open data, code, and community.

Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. The Tabula Muris Consortium (2018). *Nature*. doi:10.1038/s41586-018-0590-4

> Here we present a compendium of single-cell transcriptomic data from the model organism Mus musculus that comprises more than 100,000 cells from 20 organs and tissues. These data represent a new resource for cell biology, reveal gene expression in poorly characterized cell populations and enable the direct and controlled comparison of gene expression in cell types that are shared between tissues, such as T lymphocytes and endothelial cells from different anatomical locations. Two distinct technical approaches were used for most organs: one approach, microfluidic droplet-based 3'-end counting, enabled the survey of thousands of cells at relatively low coverage, whereas the other, full-length transcript analysis based on fluorescence-activated cell sorting, enabled the characterization of cell types with high sensitivity and coverage. The cumulative data provide the foundation for an atlas of transcriptomic cell biology.

The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. The Tabula Sapiens Consortium (2022). *Science*. doi:10.1126/science.abl4896

> Molecular characterization of cell types using single-cell transcriptome sequencing is revolutionizing cell biology and enabling new insights into the physiology of human organs. We created a human reference atlas comprising nearly 500,000 cells from 24 different tissues and organs, many from the same donor. This atlas enabled molecular characterization of more than 400 cell types, their distribution across tissues, and tissue-specific variation in gene expression. Using multiple tissues from a single donor enabled identification of the clonal distribution of T cells between tissues, identification of the tissue-specific mutation rate in B cells, and analysis of the cell cycle state and proliferative potential of shared cell types across tissues. Cell type-specific RNA splicing was discovered and analyzed across tissues within an individual.

- Zeng (2022) — what is a cell type?

  What is a cell type and how to define it?. Zeng (2022). *Cell*. doi:10.1016/j.cell.2022.06.031

  > Cell types are the basic functional units of an organism. Cell types exhibit diverse phenotypic properties at multiple levels, making them challenging to define, categorize, and understand. This review provides an overview of the basic principles of cell types rooted in evolution and development and discusses approaches to characterize and classify cell types and investigate how they contribute to the organism's function, using the mammalian brain as a primary example. I propose a roadmap toward a conceptual framework and knowledge base of cell types that will enable a deeper understanding of the dynamic changes of cellular function under healthy and diseased conditions.

B.7.5 Cancer evolution

- Nowell (1976) — the clonal evolution of tumour cell populations.

  The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression.. Nowell (1976). *Science*. doi:10.1126/science.959840

  > It is proposed that most neoplasms arise from a single cell of origin, and tumor progression results from acquired genetic variability within the original clone allowing sequential selection of more aggressive sublines. Tumor cell populations are apparently more genetically unstable than normal cells, perhaps from activation of specific gene loci in the neoplasm, continued presence of carcinogen, or even nutritional deficiencies within the tumor. The acquired genetic insta0ility and associated selection process, most readily recognized cytogenetically, results in advanced human malignancies being highly individual karyotypically and biologically. Hence, each patient's cancer may require individual specific therapy, and even this may be thwarted by emergence of a genetically variant subline resistant to the treatment. More research should be directed toward understanding and controlling the evolutionary process in tumors before it reaches the late stage usually seen in clinical cancer.

- Gerlinger et al. (2012), McGranahan and Swanton (2017) — intratumour heterogeneity and branched evolution.

  Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. Gerlinger et al. (2012). *New England Journal of Medicine*. doi:10.1056/nejmoa1113205

> Background Intratumor heterogeneity may foster tumor evolution and adaptation and hinder personalized-medicine strategies that depend on results from single tumor-biopsy samples. Methods To examine intratumor heterogeneity, we performed exome sequencing, chromosome aberration analysis, and ploidy profiling on multiple spatially separated samples obtained from primary renal carcinomas and associated metastatic sites. We characterized the consequences of intratumor heterogeneity using immunohistochemical analysis, mutation functional analysis, and profiling of messenger RNA expression. Results Phylogenetic reconstruction revealed branched evolutionary tumor growth, with 63 to 69% of all somatic mutations not detectable across every tumor region. Intratumor heterogeneity was observed for a mutation within an autoinhibitory domain of the mammalian target of rapamycin (mTOR) kinase, correlating with S6 and 4EBP phosphorylation in vivo and constitutive activation of mTOR kinase activity in vitro. Mutational intratumor heterogeneity was seen for multiple tumor-suppressor genes converging on loss of function; SETD2, PTEN, and KDM5C underwent multiple distinct and spatially separated inactivating mutations within a single tumor, suggesting convergent phenotypic evolution. Gene-expression signatures of good and poor prognosis were detected in different regions of the same tumor. Allelic composition and ploidy profiling analysis revealed extensive intratumor heterogeneity, with 26 of 30 tumor samples from four tumors harboring divergent allelic-imbalance profiles and with ploidy heterogeneity in two of four tumors. Conclusions Intratumor heterogeneity can lead to underestimation of the tumor genomics landscape portrayed from single tumor-biopsy samples and may present major challenges to personalized-medicine and biomarker development. Intratumor heterogeneity, associated with heterogeneous protein function, may foster tumor adaptation and therapeutic failure through Darwinian selection. (Funded by the Medical Research Council and others.).

Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. McGranahan & Swanton (2017). *Cell*. doi:10.1016/j.cell.2017.01.018

> Intratumor heterogeneity, which fosters tumor evolution, is a key challenge in cancer medicine. Here, we review data and technologies that have revealed intra-tumor heterogeneity across cancer types and the dynamics, constraints, and contingencies inherent to tumor evolution. We emphasize the importance of macro-evolutionary leaps, often involving large-scale chromosomal alterations, in driving tumor evolution and metastasis and consider the role of the tumor microenvironment in engendering heterogeneity and drug resistance. We suggest that bold approaches to drug development, harnessing the adaptive properties of the immune-microenvironment while limiting those of the tumor, combined with advances in clinical trial-design, will improve patient outcome.

- Schwartz and Schäffer (2017) — tumour phylogenetics.

The evolution of tumour phylogenetics: principles and practice. Schwartz & Schäffer (2017). *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

> Rapid advances in high-throughput sequencing and a growing realization of the importance of evolutionary theory to cancer genomics have led to a proliferation of phylogenetic studies of tumour progression. These studies have yielded not only new insights but also a plethora of experimental approaches, sometimes reaching conflicting or poorly supported conclusions. Here, we consider this body of work in light of the key computational principles underpinning phylogenetic inference, with the goal of providing practical guidance on the design and analysis of scientifically rigorous tumour phylogeny studies. We survey the range of methods and tools available to the researcher, their key applications, and the various unsolved problems, closing with a perspective on the prospects and broader implications of this field.

- Gawad et al. (2016) — single-cell genome sequencing.

Single-cell genome sequencing: current state of the science. Gawad et al. (2016). *Nature Reviews Genetics*. doi:10.1038/nrg.2015.16

> The field of single-cell genomics is advancing rapidly and is generating many new insights into complex biological systems, ranging from the diversity of microbial ecosystems to the genomics of human cancer. In this Review, we provide an overview of the current state of the field of single-cell genome sequencing. First, we focus on the technical challenges of making measurements that start from a single molecule of DNA, and then explore how some of these recent methodological advancements have enabled the discovery of unexpected new biology. Areas highlighted include the application of single-cell genomics to interrogate microbial dark matter and to evaluate the pathogenic roles of genetic mosaicism in multicellular organisms, with a focus on cancer. We then attempt to predict advances we expect to see in the next few years.

- Navin et al. (2011), Minussi et al. (2021) — tumour evolution from single-cell DNA sequencing.

  Tumour evolution inferred by single-cell sequencing. Navin et al. (2011). *Nature*. doi:10.1038/nature09807

  > Genomic analysis provides insights into the role of copy number variation in disease, but most methods are not designed to resolve mixed populations of cells. In tumours, where genetic heterogeneity is common, very important information may be lost that would be useful for reconstructing evolutionary history. Here we show that with flow-sorted nuclei, whole genome amplification and next generation sequencing we can accurately quantify genomic copy number within an individual nucleus. We apply single-nucleus sequencing to investigate tumour population structure and evolution in two human breast cancer cases. Analysis of 100 single cells from a polygenomic tumour revealed three distinct clonal subpopulations that probably represent sequential clonal expansions. Additional analysis of 100 single cells from a monogenomic primary tumour and its liver metastasis indicated that a single clonal expansion formed the primary tumour and seeded the metastasis. In both primary tumours, we also identified an unexpectedly abundant subpopulation of genetically diverse 'pseudodiploid' cells that do not travel to the metastatic site. In contrast to gradual models of tumour progression, our data indicate that tumours grow by punctuated clonal expansions with few persistent intermediates.

  Breast tumours maintain a reservoir of subclonal diversity during expansion. Minussi et al. (2021). *Nature*. doi:10.1038/s41586-021-03357-x

  > Our knowledge of copy number evolution during the expansion of primary breast tumours is limited 1,2 . Here, to investigate this process, we developed a single-cell, single-molecule DNA-sequencing method and performed copy number analysis of 16,178 single cells from 8 human triple-negative breast cancers and 4 cell lines. The results show that breast tumours and cell lines comprise a large milieu of subclones (7-22) that are organized into a few (3-5) major superclones. Evolutionary analysis suggests that after clonal TP53 mutations, multiple loss-of-heterozygosity events and genome doubling, there was a period of transient genomic instability followed by ongoing copy number evolution during the primary tumour expansion. By subcloning single daughter cells in culture, we show that tumour cells rediversify their genomes and do not retain isogenic properties. These data show that triple-negative breast cancers continue to evolve chromosome aberrations and maintain a reservoir of subclonal diversity during primary tumour growth.

- Patel et al. (2014), Tirosh et al. (2016) — single-cell RNA-seq of tumours.

  Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Patel et al. (2014). *Science*. doi:10.1126/science.1254257

  > Human cancers are complex ecosystems composed of cells with distinct phenotypes, genotypes, and epigenetic states, but current models do not adequately reflect tumor composition in patients. We used single-cell RNA sequencing (RNA-seq) to profile 430 cells from five primary glioblastomas, which we found to be inherently variable in their expression of diverse transcriptional programs related to oncogenic signaling, proliferation, complement/immune response, and hypoxia. We also observed a continuum of stemness-related expression states that enabled us to identify putative regulators of stemness in vivo. Finally, we show that established glioblastoma subtype classifiers are variably expressed across individual cells within a tumor and demonstrate the potential prognostic implications of such intratumoral heterogeneity. Thus, we reveal previously unappreciated heterogeneity in diverse regulatory programs central to glioblastoma biology, prognosis, and therapy.

  Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Tirosh et al. (2016). *Science*. doi:10.1126/science.aad0501

  > To explore the distinct genotypic and phenotypic states of melanoma tumors, we applied single-cell RNA sequencing (RNA-seq) to 4645 single cells isolated from 19 patients, profiling malignant, immune, stromal, and endothelial cells. Malignant cells within the same tumor displayed transcriptional heterogeneity associated with the cell cycle, spatial context, and a drug-resistance program. In particular, all tumors harbored malignant cells from two distinct transcriptional cell states, such that tumors characterized by high levels of the MITF transcription factor also contained cells with low MITF and elevated levels of the AXL kinase. Single-cell analyses suggested distinct tumor microenvironmental patterns, including cell-to-cell interactions. Analysis of tumor-infiltrating T cells revealed exhaustion programs, their connection to T cell activation and clonal expansion, and their variability across patients. Overall, we begin to unravel

the cellular ecosystem of tumors and how single-cell genomics offers insights with implications for both targeted and immune therapies.

- Gao et al. (2021) — copy number from single-cell transcriptomes.

  Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Gao et al. (2021). *Nature Biotechnology*. doi:10.1038/s41587-020-00795-2

  > Single-cell transcriptomic analysis is widely used to study human tumors. However, it remains challenging to distinguish normal cell types in the tumor microenvironment from malignant cells and to resolve clonal substructure within the tumor. To address these challenges, we developed an integrative Bayesian segmentation approach called copy number karyotyping of aneuploid tumors (CopyKAT) to estimate genomic copy number profiles at an average genomic resolution of 5 Mb from read depth in high-throughput single-cell RNA sequencing (scRNA-seq) data. We applied CopyKAT to analyze 46,501 single cells from 21 tumors, including triple-negative breast cancer, pancreatic ductal adenocarcinoma, anaplastic thyroid cancer, invasive ductal carcinoma and glioblastoma, to accurately (98%) distinguish cancer cells from normal cell types. In three breast tumors, CopyKAT resolved clonal subpopulations that differed in the expression of cancer genes, such as KRAS, and signatures, including epithelial-to-mesenchymal transition, DNA repair, apoptosis and hypoxia. These data show that CopyKAT can aid in the analysis of scRNA-seq data in a variety of solid human tumors.

- Lawson et al. (2018) — tumour heterogeneity and metastasis at single-cell resolution.

  Tumour heterogeneity and metastasis at single-cell resolution. Lawson et al. (2018). *Nature Cell Biology*. doi:10.1038/s41556-018-0236-7

  > Tumours comprise a heterogeneous collection of cells with distinct genetic and phenotypic properties that can differentially promote progression, metastasis and drug resistance. Emerging single-cell technologies provide a new opportunity to profile individual cells within tumours and investigate what roles they play in these processes. This Review discusses key technological considerations for single-cell studies in cancer, new findings using single-cell technologies and critical open questions for future applications.

B.8 Proteomics

Further papers for Chapter 8.

In the HTML version, click *Abstract* to show the abstract.

B.8.1 Reviews

- Aebersold and Mann (2003) — mass spectrometry-based proteomics.

  Mass spectrometry-based proteomics. Aebersold & Mann (2003). *Nature*. doi:10.1038/nature01511

  > Recent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology. These include the study of protein-protein interactions via affinity-based isolations on a small and proteome-wide scale, the mapping of numerous organelles, the concurrent description of the malaria parasite genome and proteome, and the generation of quantitative protein profiles from diverse species. The ability of mass spectrometry to identify and, increasingly, to precisely quantify thousands of proteins from complex samples can be expected to impact broadly on biology and medicine.

- Nesvizhskii (2010) — computational methods and error-rate estimation for peptide and protein identification.

  A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Nesvizhskii (2010). *Journal of Proteomics*. doi:10.1016/j.jprot.2010.08.009

> This manuscript provides a comprehensive review of the peptide and protein identification process using tandem mass spectrometry (MS/MS) data generated in shotgun proteomic experiments. The commonly used methods for assigning peptide sequences to MS/MS spectra are critically discussed and compared, from basic strategies to advanced multi-stage approaches. A particular attention is paid to the problem of false-positive identifications. Existing statistical approaches for assessing the significance of peptide to spectrum matches are surveyed, ranging from single-spectrum approaches such as expectation values to global error rate estimation procedures such as false discovery rates and posterior probabilities. The importance of using auxiliary discriminant information (mass accuracy, peptide separation coordinates, digestion properties, and etc.) is discussed, and advanced computational approaches for joint modeling of multiple sources of information are presented. This review also includes a detailed analysis of the issues affecting the interpretation of data at the protein level, including the amplification of error rates when going from peptide to protein level, and the ambiguities in inferring the identifies of sample proteins in the presence of shared peptides. Commonly used methods for computing protein-level confidence scores are discussed in detail. The review concludes with a discussion of several outstanding computational issues.

- Toby et al. (2016) — top-down proteomics and proteoforms.

  Progress in Top-Down Proteomics and the Analysis of Proteoforms. Toby et al. (2016). *Annual Review of Analytical Chemistry*. doi:10.1146/annurev-anchem-071015-041550

  > From a molecular perspective, enactors of function in biology are intact proteins that can be variably modified at the genetic, transcriptional, or post-translational level. Over the past 30 years, mass spectrometry (MS) has become a powerful method for the analysis of proteomes. Prevailing bottom-up proteomics operates at the level of the peptide, leading to issues with protein inference, connectivity, and incomplete sequence/modification information. Top-down proteomics (TDP), alternatively, applies MS at the proteoform level to analyze intact proteins with diverse sources of intramolecular complexity preserved during analysis. Fortunately, advances in prefractionation workflows, MS instrumentation, and dissociation methods for whole-protein ions have helped TDP emerge as an accessible and potentially disruptive modality with increasingly translational value. In this review, we discuss technical and conceptual advances in TDP, along with the growing power of proteoform-resolved measurements in clinical and translational research.

B.8.2 History and instruments

- O'Farrell (1975) — two-dimensional gel electrophoresis.

  High resolution two-dimensional electrophoresis of proteins.. O'Farrell (1975). *Journal of Biological Chemistry*. doi:10.1016/s0021-9258(19)41496-8

  > A technique has been developed for the separation of proteins by two-dimensional polyacrylamide gel electrophoresis. Due to its resolution and sensitivity, this technique is a powerful tool for the analysis and detection of proteins from complex biological sources. Proteins are separated according to isoelectric point by isoelectric focusing in the first dimension, and according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since these two parameters are unrelated, it is possible to obtain an almost uniform distribution of protein spots across a two-diminsional gel. This technique has resolved 1100 different components from Escherichia coli and should be capable of resolving a maximum of 5000 proteins. A protein containing as little as one disintegration per min of either 14C or 35S can be detected by autoradiography. A protein which constitutes 10 minus 4 to 10 minus 5% of the total protein can be detected and quantified by autoradiography. The reproducibility of the separation is sufficient to permit each spot on one separation to be matched with a spot on a different separation. This technique provides a method for estimation (at the described sensitivities) of the number of proteins made by any biological system. This system can resolve proteins differing in a single charge and consequently can be used in the analysis of in vivo modifications resulting in a change in charge. Proteins whose charge is changed by missense mutations can be identified. A detailed description of the methods as well as the characteristics of this system are presented.

- Fenn et al. (1989), Karas and Hillenkamp (1988), Tanaka et al. (1988) — electrospray and MALDI.

  Electrospray Ionization for Mass Spectrometry of Large Biomolecules. Fenn et al. (1989). *Science*. doi:10.1126/science.2675315

> Electrospray ionization has recently emerged as a powerful technique for producing intact ions in vacuo from large and complex species in solution. To an extent greater than has previously been possible with the more familiar "soft" ionization methods, this technique makes the power and elegance of mass spectrometric analysis applicable to the large and fragile polar molecules that play such vital roles in biological systems. The distinguishing features of electrospray spectra for large molecules are coherent sequences of peaks whose component ions are multiply charged, the ions of each peak differing by one charge from those of adjacent neighbors in the sequence. Spectra have been obtained for biopolymers including oligonucleotides and proteins, the latter having molecular weights up to 130,000, with as yet no evidence of an upper limit.

Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Karas & Hillenkamp (1988). *Analytical Chemistry*. doi:10.1021/ac00171a028

Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Tanaka et al. (1988). *Rapid Communications in Mass Spectrometry*. doi:10.1002/rcm.1290020802

- Henzel et al. (1993) — peptide mass fingerprinting.

Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.. Henzel et al. (1993). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.90.11.5011

> A rapid method for the identification of known proteins separated by two-dimensional gel electrophoresis is described in which molecular masses of peptide fragments are used to search a protein sequence database. The peptides are generated by in situ reduction, alkylation, and tryptic digestion of proteins electroblotted from two-dimensional gels. Masses are determined at the subpicomole level by matrix-assisted laser desorption/ionization mass spectrometry of the unfractionated digest. A computer program has been developed that searches the protein sequence database for multiple peptides of individual proteins that match the measured masses. To ensure that the most recent database updates are included, a theoretical digest of the entire database is generated each time the program is executed. This method facilitates simultaneous processing of a large number of two-dimensional gel spots. The method was applied to a two-dimensional gel of a crude Escherichia coli extract that was electroblotted onto poly(vinylidene difluoride) membrane. Ten randomly chosen spots were analyzed. With as few as three peptide masses, each protein was uniquely identified from over 91,000 protein sequences. All identifications were verified by concurrent N-terminal sequencing of identical spots from a second blot. One of the spots contained an N-terminally blocked protein that required enzymatic cleavage, peptide separation, and Edman degradation for confirmation of its identity.

- Makarov (2000), Hu et al. (2005) — the Orbitrap.

Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Makarov (2000). *Analytical Chemistry*. doi:10.1021/ac991131p

> This work describes a new type of mass analyzer which employs trapping in an electrostatic field. The potential distribution of the field can be represented as a combination of quadrupole and logarithmic potentials. In the absence of any magnetic or rf fields, ion stability is achieved only due to ions orbiting around an axial electrode. Orbiting ions also perform harmonic oscillations along the electrode with frequency proportional to (m/z)-1/2. These oscillations are detected using image current detection and are transformed into mass spectra using fast FT, similarly to FTICR. Practical aspects of the trap design are presented. High-mass resolution up to 150,000 for ions produced by laser ablation has been demonstrated, along with high-energy acceptance and wide mass range.

The Orbitrap: a new mass spectrometer. Hu et al. (2005). *Journal of Mass Spectrometry*. doi:10.1002/jms.856

> Research areas such as proteomics and metabolomics are driving the demand for mass spectrometers that have high performance but modest power requirements, size, and cost. This paper describes such an instrument, the Orbitrap, based on a new type of mass analyzer invented by Makarov. The Orbitrap operates by radially trapping ions about a central spindle electrode. An outer barrel-like electrode is coaxial with the inner spindlelike electrode and mass/charge values are measured from the frequency of harmonic ion oscillations, along the axis of the electric field, undergone by the orbitally trapped ions.

This axial frequency is independent of the energy and spatial spread of the ions. Ion frequencies are measured non-destructively by acquisition of time-domain image current transients, with subsequent fast Fourier transforms (FFTs) being used to obtain the mass spectra. In addition to describing the Orbitrap mass analyzer, this paper also describes a complete Orbitrap-based mass spectrometer, equipped with an electrospray ionization source (ESI). Ions are transferred from the ESI source through three stages of differential pumping using RF guide quadrupoles. The third quadrupole, pressurized to less than 10(-3) Torr with collision gas, acts as an ion accumulator; ion/neutral collisions slow the ions and cause them to pool in an axial potential well at the end of the quadrupole. Ion bunches are injected from this pool into the Orbitrap analyzer for mass analysis. The ion injection process is described in a simplified way, including a description of electrodynamic squeezing, field compensation for the effects of the ion injection slit, and criteria for orbital stability. Features of the Orbitrap at its present stage of development include high mass resolution (up to 150,000), large space charge capacity, high mass accuracy (2-5 ppm), a mass/charge range of at least 6000, and dynamic range greater than 10(3). Applications based on electrospray ionization are described, including characterization of transition-metal complexes, oligosaccharides, peptides, and proteins. Use is also made of the high-resolution capabilities of the Orbitrap to confirm the presence of metaclusters of serine octamers in ESI mass spectra and to perform H/D exchange experiments on these ions in the storage quadrupole.

- Meier et al. (2018) — PASEF on a trapped ion mobility instrument.

  Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. Meier et al. (2018). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.tir118.000900

  In bottom-up proteomics, peptides are separated by liquid chromatography with elution peak widths in the range of seconds, whereas mass spectra are acquired in about 100 microseconds with time-of-flight (TOF) instruments. This allows adding ion mobility as a third dimension of separation. Among several formats, trapped ion mobility spectrometry (TIMS) is attractive because of its small size, low voltage requirements and high efficiency of ion utilization. We have recently demonstrated a scan mode termed parallel accumulation - serial fragmentation (PASEF), which multiplies the sequencing speed without any loss in sensitivity (Meier et al. , PMID: 26538118). Here we introduce the timsTOF Pro instrument, which optimally implements online PASEF. It features an orthogonal ion path into the ion mobility device, limiting the amount of debris entering the instrument and making it very robust in daily operation. We investigate different precursor selection schemes for shotgun proteomics to optimally allocate in excess of 100 fragmentation events per second. More than 600,000 fragmentation spectra in standard 120 min LC runs are achievable, which can be used for near exhaustive precursor selection in complex mixtures or accumulating the signal of weak precursors. In 120 min single runs of HeLa digest, MaxQuant identified more than 6,000 proteins without matching to a library and with high quantitative reproducibility (R > 0.97). Online PASEF achieves a remarkable sensitivity with more than 2,500 proteins identified in 30 min runs of only 10 ng HeLa digest. We also show that highly reproducible collisional cross sections can be acquired on a large scale (R > 0.99). PASEF on the timsTOF Pro is a valuable addition to the technological toolbox in proteomics, with a number of unique operating modes that are only beginning to be explored.

B.8.3 Acquisition and quantification

- Gillet et al. (2012) — SWATH-MS.

  Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Gillet et al. (2012). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.o111.016717

  Most proteomic studies use liquid chromatography coupled to tandem mass spectrometry to identify and quantify the peptides generated by the proteolysis of a biological sample. However, with the current methods it remains challenging to rapidly, consistently, reproducibly, accurately, and sensitively detect and quantify large fractions of proteomes across multiple samples. Here we present a new strategy that systematically queries sample sets for the presence and quantity of essentially any protein of interest. It consists of using the information available in fragment ion spectral libraries to mine the complete fragment ion maps generated using a data-independent acquisition method. For this study, the data were acquired on a fast, high resolution quadrupole-quadrupole time-of-flight (TOF) instrument by repeatedly cycling through 32 consecutive 25-Da precursor isolation windows (swaths). This SWATH MS acquisition setup generates, in a single sample injection, time-resolved fragment ion spectra for all the analytes

detectable within the 400-1200 m/z precursor range and the user-defined retention time window. We show that suitable combinations of fragment ions extracted from these data sets are sufficiently specific to confidently identify query peptides over a dynamic range of 4 orders of magnitude, even if the precursors of the queried peptides are not detectable in the survey scans. We also show that queried peptides are quantified with a consistency and accuracy comparable with that of selected reaction monitoring, the gold standard proteomic quantification method. Moreover, targeted data extraction enables ad libitum quantification refinement and dynamic extension of protein probing by iterative re-mining of the once-and-forever acquired data sets. This combination of unbiased, broad range precursor ion fragmentation and targeted data extraction alleviates most constraints of present proteomic methods and should be equally applicable to the comprehensive analysis of other classes of analytes, beyond proteomics.

- Lange et al. (2008), Peterson et al. (2012) — targeted proteomics (SRM and PRM).

Selected reaction monitoring for quantitative proteomics: a tutorial. Lange et al. (2008). *Molecular Systems Biology*. doi:10.1038/msb.2008.61

> Systems biology relies on data sets in which the same group of proteins is consistently identified and precisely quantified across multiple samples, a requirement that is only partially achieved by current proteomics approaches. Selected reaction monitoring (SRM)-also called multiple reaction monitoring-is emerging as a technology that ideally complements the discovery capabilities of shotgun strategies by its unique potential for reliable quantification of analytes of low abundance in complex mixtures. In an SRM experiment, a predefined precursor ion and one of its fragments are selected by the two mass filters of a triple quadrupole instrument and monitored over time for precise quantification. A series of transitions (precursor/fragment ion pairs) in combination with the retention time of the targeted peptide can constitute a definitive assay. Typically, a large number of peptides are quantified during a single LC-MS experiment. This tutorial explains the application of SRM for quantitative proteomics, including the selection of proteotypic peptides and the optimization and validation of transitions. Furthermore, normalization and various factors affecting sensitivity and accuracy are discussed.

Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. Peterson et al. (2012). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.o112.020131

> Selected reaction monitoring on a triple quadrupole mass spectrometer is currently experiencing a renaissance within the proteomics community for its, as yet, unparalleled ability to characterize and quantify a set of proteins reproducibly, completely, and with high sensitivity. Given the immense benefit that high resolution and accurate mass instruments have brought to the discovery proteomics field, we wondered if highly accurate mass measurement capabilities could be leveraged to provide benefits in the targeted proteomics domain as well. Here, we propose a new targeted proteomics paradigm centered on the use of next generation, quadrupole-equipped high resolution and accurate mass instruments: parallel reaction monitoring (PRM). In PRM, the third quadrupole of a triple quadrupole is substituted with a high resolution and accurate mass mass analyzer to permit the parallel detection of all target product ions in one, concerted high resolution mass analysis. We detail the analytical performance of the PRM method, using a quadrupole-equipped bench-top Orbitrap MS, and draw a performance comparison to selected reaction monitoring in terms of run-to-run reproducibility, dynamic range, and measurement accuracy. In addition to requiring minimal upfront method development and facilitating automated data analysis, PRM yielded quantitative data over a wider dynamic range than selected reaction monitoring in the presence of a yeast background matrix because of PRM's high selectivity in the mass-to-charge domain. With achievable linearity over the quantifiable dynamic range found to be statistically equal between the two methods, our investigation suggests that PRM will be a promising new addition to the quantitative proteomics toolbox.

- Ong et al. (2002) — SILAC.

Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Ong et al. (2002). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.m200025-mcp200

> Quantitative proteomics has traditionally been performed by two-dimensional gel electrophoresis, but recently, mass spectrometric methods based on stable isotope quantitation have shown great promise for the simultaneous and automated identification and quantitation of complex protein mixtures. Here we describe a method, termed SILAC, for stable isotope labeling by amino acids in cell culture, for the in vivo incorporation of specific amino acids into all mammalian proteins. Mammalian cell lines are grown

in media lacking a standard essential amino acid but supplemented with a non-radioactive, isotopically labeled form of that amino acid, in this case deuterated leucine (Leu-d3). We find that growth of cells maintained in these media is no different from growth in normal media as evidenced by cell morphology, doubling time, and ability to differentiate. Complete incorporation of Leu-d3 occurred after five doublings in the cell lines and proteins studied. Protein populations from experimental and control samples are mixed directly after harvesting, and mass spectrometric identification is straightforward as every leucine-containing peptide incorporates either all normal leucine or all Leu-d3. We have applied this technique to the relative quantitation of changes in protein expression during the process of muscle cell differentiation. Proteins that were found to be up-regulated during this process include glyceraldehyde-3-phosphate dehydrogenase, fibronectin, and pyruvate kinase M2. SILAC is a simple, inexpensive, and accurate procedure that can be used as a quantitative proteomic approach in any cell culture system.

- Ross et al. (2004), Thompson et al. (2003) — iTRAQ and TMT.

Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Ross et al. (2004). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.m400129-mcp200

> We describe here a multiplexed protein quantitation strategy that provides relative and absolute measurements of proteins in complex mixtures. At the core of this methodology is a multiplexed set of isobaric reagents that yield amine-derivatized peptides. The derivatized peptides are indistinguishable in MS, but exhibit intense low-mass MS/MS signature ions that support quantitation. In this study, we have examined the global protein expression of a wild-type yeast strain and the isogenic upf1Delta and xrn1Delta mutant strains that are defective in the nonsense-mediated mRNA decay and the general 5' to 3' decay pathways, respectively. We also demonstrate the use of 4-fold multiplexing to enable relative protein measurements simultaneously with determination of absolute levels of a target protein using synthetic isobaric peptide standards. We find that inactivation of Upf1p and Xrn1p causes common as well as unique effects on protein expression.

Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Thompson et al. (2003). *Analytical Chemistry*. doi:10.1021/ac0262560

> A novel MS/MS-based analysis strategy using isotopomer labels, referred to as "tandem mass tags" (TMTs), for the accurate quantification of peptides and proteins is described. The new tags are designed to ensure that identical peptides labeled with different TMTs exactly comigrate in all separations. The tags require novel methods of quantification analysis using tandem mass spectrometry. The new tags and analysis methods allow peptides from different samples to be identified by their relative abundance with greater ease and accuracy than other methods. The new TMTs permit simultaneous determination of both the identity and relative abundances of peptide pairs using a collision induced dissociation (CID)-based analysis method. Relative abundance measurements made in the MS/MS mode using the new tags are accurate and sensitive. Compared to MS-mode measurements, a very high signal-to-noise ratio is achieved with MS/MS based detection. The new tags should be applicable to a wide variety of peptide isolation methods.

- Cox et al. (2014) — MaxLFQ label-free quantification.

Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Cox et al. (2014). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.m113.031591

> Protein quantification without isotopic labels has been a long-standing interest in the proteomics field. However, accurate and robust proteome-wide quantification with label-free approaches remains a challenge. We developed a new intensity determination and normalization procedure called MaxLFQ that is fully compatible with any peptide or protein separation prior to LC-MS analysis. Protein abundance profiles are assembled using the maximum possible information from MS signals, given that the presence of quantifiable peptides varies from sample to sample. For a benchmark dataset with two proteomes mixed at known ratios, we accurately detected the mixing ratio over the entire protein expression range, with greater precision for abundant proteins. The significance of individual label-free quantifications was obtained via a t test approach. For a second benchmark dataset, we accurately quantify fold changes over several orders of magnitude, a task that is challenging with label-based methods. MaxLFQ is a generic label-free quantification technology that is readily applicable to many biological questions; it is

compatible with standard statistical analysis workflows, and it has been validated in many and diverse biological projects. Our algorithms can handle very large experiments of 500+ samples in a manageable computing time. It is implemented in the freely available MaxQuant computational proteomics platform and works completely seamlessly at the click of a button.

- Gerber et al. (2003) — absolute quantification with heavy peptides.

  Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Gerber et al. (2003). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0832254100

  A need exists for technologies that permit the direct quantification of differences in protein and posttranslationally modified protein expression levels. Here we present a strategy for the absolute quantification (termed AQUA) of proteins and their modification states. Peptides are synthesized with incorporated stable isotopes as ideal internal standards to mimic native peptides formed by proteolysis. These synthetic peptides can also be prepared with covalent modifications (e.g., phosphorylation, methylation, acetylation, etc.) that are chemically identical to naturally occurring posttranslational modifications. Such AQUA internal standard peptides are then used to precisely and quantitatively measure the absolute levels of proteins and posttranslationally modified proteins after proteolysis by using a selected reaction monitoring analysis in a tandem mass spectrometer. In the present work, the AQUA strategy was used to (i) quantify low abundance yeast proteins involved in gene silencing, (ii) quantitatively determine the cell cycle-dependent phosphorylation of Ser-1126 of human separase protein, and (iii) identify kinases capable of phosphorylating Ser-1501 of separase in an in vitro kinase assay. The methods described here represent focused, alternative approaches for studying the dynamically changing proteome.

B.8.4 Identification algorithms

- Eng et al. (1994) — SEQUEST.

  An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Eng et al. (1994). *Journal of the American Society for Mass Spectrometry*. doi:10.1016/1044-0305(94)80016-2

  A method to correlate the uninterpreted tandem mass spectra of peptides produced under low energy (10–50 eV) collision conditions with amino acid sequences in the Genpept database has been developed. In this method the protein database is searched to identify linear amino acid sequences within a mass tolerance of +/- 1 mu of the precursor ion molecular weight A cross-correlation function is then used to provide a measurement of similarity between the mass-to-charge ratios for the fragment ions predicted from amino acid sequences obtained from the database and the fragment ions observed in the tandem mass spectrum. In general, a difference greater than 0.1 between the normalized cross-correlation functions of the first- and second-ranked search results indicates a successful match between sequence and spectrum. Searches of species-specific protein databases with tandem mass spectra acquired from peptides obtained from the enzymatically digested total proteins of E. coli and S. cerevisiae cells allowed matching of the spectra to amino acid sequences within proteins of these organisms. The approach described in this manuscript provides a convenient method to interpret tandem mass spectra with known sequences in a protein database.

- Perkins et al. (1999) — Mascot.

  Probability-based protein identification by searching sequence databases using mass spectrometry data. Perkins et al. (1999). *Electrophoresis*. doi:10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2

  Several algorithms have been described in the literature for protein identification by searching a sequence database using mass spectrometry data. In some approaches, the experimental data are peptide molecular weights from the digestion of a protein by an enzyme. Other approaches use tandem mass spectrometry (MS/MS) data from one or more peptides. Still others combine mass data with amino acid sequence data. We present results from a new computer program, Mascot, which integrates all three types of search. The scoring algorithm is probability based, which has a number of advantages: (i) A simple rule can be used to judge whether a result is significant or not. This is particularly useful in guarding against false positives. (ii) Scores can be compared with those from other types of search, such as sequence homology. (iii) Search parameters can be readily optimised by iteration. The strengths and limitations of

probability-based scoring are discussed, particularly in the context of high throughput, fully automated protein identification.

- Craig and Beavis (2004), Cox et al. (2011), Kim and Pevzner (2014) — X!Tandem, Andromeda and MS-GF+.

TANDEM: matching proteins with tandem mass spectra. Craig & Beavis (2004). *Bioinformatics*. doi:10.1093/bioinformatics/bth092

> Tandem mass spectra obtained from fragmenting peptide ions contain some peptide sequence specific information, but often there is not enough information to sequence the original peptide completely. Several proprietary software applications have been developed to attempt to match the spectra with a list of protein sequences that may contain the sequence of the peptide. The application TANDEM was written to provide the proteomics research community with a set of components that can be used to test new methods and algorithms for performing this type of sequence-to-data matching. Availability The source code and binaries for this software are available at http://www.proteome.ca/opensource.html, for Windows, Linux and Macintosh OSX. The source code is made available under the Artistic License, from the authors.

Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. Cox et al. (2011). *Journal of Proteome Research*. doi:10.1021/pr101065j

> A key step in mass spectrometry (MS)-based proteomics is the identification of peptides in sequence databases by their fragmentation spectra. Here we describe Andromeda, a novel peptide search engine using a probabilistic scoring model. On proteome data, Andromeda performs as well as Mascot, a widely used commercial search engine, as judged by sensitivity and specificity analysis based on target decoy searches. Furthermore, it can handle data with arbitrarily high fragment mass accuracy, is able to assign and score complex patterns of post-translational modifications, such as highly phosphorylated peptides, and accommodates extremely large databases. The algorithms of Andromeda are provided. Andromeda can function independently or as an integrated search engine of the widely used MaxQuant computational proteomics platform and both are freely available at www.maxquant.org. The combination enables analysis of large data sets in a simple analysis workflow on a desktop computer. For searching individual spectra Andromeda is also accessible via a web server. We demonstrate the flexibility of the system by implementing the capability to identify cofragmented peptides, significantly improving the total number of identified peptides.

MS-GF+ makes progress towards a universal database search tool for proteomics. Kim & Pevzner (2014). *Nature Communications*. doi:10.1038/ncomms6277

> Mass spectrometry (MS) instruments and experimental protocols are rapidly advancing, but the software tools to analyse tandem mass spectra are lagging behind. We present a database search tool MS-GF+ that is sensitive (it identifies more peptides than most other database search tools) and universal (it works well for diverse types of spectra, different configurations of MS instruments and different experimental protocols). We benchmark MS-GF+ using diverse spectral data sets: (i) spectra of varying fragmentation methods; (ii) spectra of multiple enzyme digests; (iii) spectra of phosphorylated peptides; and (iv) spectra of peptides with unusual fragmentation propensities produced by a novel alpha-lytic protease. For all these data sets, MS-GF+ significantly increases the number of identified peptides compared with commonly used methods for peptide identifications. We emphasize that although MS-GF+ is not specifically designed for any particular experimental set-up, it improves on the performance of tools specifically designed for these applications (for example, specialized tools for phosphoproteomics).

- Kong et al. (2017) — MSFragger and open search.

MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Kong et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4256

> There is a need to better understand and handle the 'dark matter' of proteomics-the vast diversity of post-translational and chemical modifications that are unaccounted in a typical mass spectrometry-based analysis and thus remain unidentified. We present a fragment-ion indexing method, and its implementation in peptide identification tool MSFragger, that enables a more than 100-fold improvement in speed over most existing proteome database search tools. Using several large proteomic data sets, we demonstrate how MSFragger empowers the open database search concept for comprehensive identification of peptides and all their modified forms, uncovering dramatic differences in modification rates across experimental

samples and conditions. We further illustrate its utility using protein-RNA cross-linked peptide data and using affinity purification experiments where we observe, on average, a 300% increase in the number of identified spectra for enriched proteins. We also discuss the benefits of open searching for improved false discovery rate estimation in proteomics.

- Lam et al. (2007) — spectral library search.

  Development and validation of a spectral library searching method for peptide identification from MS/MS. Lam et al. (2007). *PROTEOMICS*. doi:10.1002/pmic.200600625

  > A notable inefficiency of shotgun proteomics experiments is the repeated rediscovery of the same identifiable peptides by sequence database searching methods, which often are time-consuming and error-prone. A more precise and efficient method, in which previously observed and identified peptide MS/MS spectra are catalogued and condensed into searchable spectral libraries to allow new identifications by spectral matching, is seen as a promising alternative. To that end, an open-source, functionally complete, high-throughput and readily extensible MS/MS spectral searching tool, SpectraST, was developed. A high-quality spectral library was constructed by combining the high-confidence identifications of millions of spectra taken from various data repositories and searched using four sequence search engines. The resulting library consists of over 30,000 spectra for Saccharomyces cerevisiae. Using this library, SpectraST vastly outperforms the sequence search engine SEQUEST in terms of speed and the ability to discriminate good and bad hits. A unique advantage of SpectraST is its full integration into the popular Trans Proteomic Pipeline suite of software, which facilitates user adoption and provides important functionalities such as peptide and protein probability assignment, quantification, and data visualization. This method of spectral library searching is especially suited for targeted proteomics applications, offering superior performance to traditional sequence searching.

- Dančík et al. (1999), Ma et al. (2003) — de novo sequencing and PEAKS.

  De Novo Peptide Sequencing via Tandem Mass Spectrometry. Dančík et al. (1999). *Journal of Computational Biology*. doi:10.1089/106652799318300

  > Peptide sequencing via tandem mass spectrometry (MS/MS) is one of the most powerful tools in proteomics for identifying proteins. Because complete genome sequences are accumulating rapidly, the recent trend in interpretation of MS/MS spectra has been database search. However, de novo MS/MS spectral interpretation remains an open problem typically involving manual interpretation by expert mass spectrometrists. We have developed a new algorithm, SHERENGA, for de novo interpretation that automatically learns fragment ion types and intensity thresholds from a collection of test spectra generated from any type of mass spectrometer. The test data are used to construct optimal path scoring in the graph representations of MS/MS spectra. A ranked list of high scoring paths corresponds to potential peptide sequences. SHERENGA is most useful for interpreting sequences of peptides resulting from unknown proteins and for validating the results of database search algorithms in fully automated, high-throughput peptide sequencing.

  PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Ma et al. (2003). *Rapid Communications in Mass Spectrometry*. doi:10.1002/rcm.1196

  > A number of different approaches have been described to identify proteins from tandem mass spectrometry (MS/MS) data. The most common approaches rely on the available databases to match experimental MS/MS data. These methods suffer from several drawbacks and cannot be used for the identification of proteins from unknown genomes. In this communication, we describe a new de novo sequencing software package, PEAKS, to extract amino acid sequence information without the use of databases. PEAKS uses a new model and a new algorithm to efficiently compute the best peptide sequences whose fragment ions can best interpret the peaks in the MS/MS spectrum. The output of the software gives amino acid sequences with confidence scores for the entire sequences, as well as an additional novel positional scoring scheme for portions of the sequences. The performance of PEAKS is compared with Lutefisk, a well-known de novo sequencing software, using quadrupole-time-of-flight (Q-TOF) data obtained for several tryptic peptides from standard proteins.

- Tran et al. (2017), Eloff et al. (2025) — DeepNovo and InstaNovo.

  De novo peptide sequencing by deep learning. Tran et al. (2017). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1705691114

> De novo peptide sequencing from tandem MS data is the key technology in proteomics for the characterization of proteins, especially for new sequences, such as mAbs. In this study, we propose a deep neural network model, DeepNovo, for de novo peptide sequencing. DeepNovo architecture combines recent advances in convolutional neural networks and recurrent neural networks to learn features of tandem mass spectra, fragment ions, and sequence patterns of peptides. The networks are further integrated with local dynamic programming to solve the complex optimization task of de novo sequencing. We evaluated the method on a wide variety of species and found that DeepNovo considerably outperformed state of the art methods, achieving 7.7-22.9% higher accuracy at the amino acid level and 38.1-64.0% higher accuracy at the peptide level. We further used DeepNovo to automatically reconstruct the complete sequences of antibody light and heavy chains of mouse, achieving 97.5-100% coverage and 97.2-99.5% accuracy, without assisting databases. Moreover, DeepNovo is retrainable to adapt to any sources of data and provides a complete end-to-end training and prediction solution to the de novo sequencing problem. Not only does our study extend the deep learning revolution to a new field, but it also shows an innovative approach in solving optimization problems by using deep learning and dynamic programming.

InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. Eloff et al. (2025). *Nature Machine Intelligence*. doi:10.1038/s42256-025-01019-5

> Mass spectrometry-based proteomics focuses on identifying the peptide that generates a tandem mass spectrum. Traditional methods rely on protein databases but are often limited or inapplicable in certain contexts. De novo peptide sequencing, which assigns peptide sequences to spectra without prior information, is valuable for diverse biological applications; however, owing to a lack of accuracy, it remains challenging to apply. Here we introduce InstaNovo, a transformer model that translates fragment ion peaks into peptide sequences. We demonstrate that InstaNovo outperforms state-of-the-art methods and showcase its utility in several applications. We also introduce InstaNovo+, a diffusion model that improves performance through iterative refinement of predicted sequences. Using these models, we achieve improved therapeutic sequencing coverage, discover novel peptides and detect unreported organisms in diverse datasets, thereby expanding the scope and detection rate of proteomics searches. Our models unlock opportunities across domains such as direct protein sequencing, immunopeptidomics and exploration of the dark proteome.

- Gessulat et al. (2019), Gabriels et al. (2019), Zeng et al. (2022) — Prosit, MS²PIP and AlphaPeptDeep.

Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Gessulat et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0426-7

> In mass-spectrometry-based proteomics, the identification and quantification of peptides and proteins heavily rely on sequence database searching or spectral library matching. The lack of accurate predictive models for fragment ion intensities impairs the realization of the full potential of these approaches. Here, we extended the ProteomeTools synthetic peptide library to 550,000 tryptic peptides and 21 million high-quality tandem mass spectra. We trained a deep neural network, termed Prosit, resulting in chromatographic retention time and fragment ion intensity predictions that exceed the quality of the experimental data. Integrating Prosit into database search pipelines led to more identifications at >10× lower false discovery rates. We show the general applicability of Prosit by predicting spectra for proteases other than trypsin, generating spectral libraries for data-independent acquisition and improving the analysis of metaproteomes. Prosit is integrated into ProteomicsDB, allowing search result re-scoring and custom spectral library generation for any organism on the basis of peptide sequence alone.

Updated MS²PIP web server delivers fast and accurate MS² peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Gabriels et al. (2019). *Nucleic Acids Research*. doi:10.1093/nar/gkz299

> MS²PIP is a data-driven tool that accurately predicts peak intensities for a given peptide's fragmentation mass spectrum. Since the release of the MS²PIP web server in 2015, we have brought significant updates to both the tool and the web server. In addition to the original models for CID and HCD fragmentation, we have added specialized models for the TripleTOF 5600+ mass spectrometer, for TMT-labeled peptides, for iTRAQ-labeled peptides, and for iTRAQ-labeled phosphopeptides. Because the fragmentation pattern is heavily altered in each of these cases, these additional models greatly improve the prediction accuracy for their corresponding data types. We have also substantially reduced the computational resources required to run MS²PIP, and have completely rebuilt the web server, which now allows predictions of up to 100 000 peptide sequences in a single request. The MS²PIP web server is freely available at https://iomics.ugent.be/ms2pip/.

AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. Zeng et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-34904-3

> Machine learning and in particular deep learning (DL) are increasingly important in mass spectrometry (MS)-based proteomics. Recent DL models can predict the retention time, ion mobility and fragment intensities of a peptide just from the amino acid sequence with good accuracy. However, DL is a very rapidly developing field with new neural network architectures frequently appearing, which are challenging to incorporate for proteomics researchers. Here we introduce AlphaPeptDeep, a modular Python framework built on the PyTorch DL library that learns and predicts the properties of peptides (https://github.com/MannLabs/alphapeptdeep). It features a model shop that enables non-specialists to create models in just a few lines of code. AlphaPeptDeep represents post-translational modifications in a generic manner, even if only the chemical composition is known. Extensive use of transfer learning obviates the need for large data sets to refine models for particular experimental conditions. The AlphaPeptDeep models for predicting retention time, collisional cross sections and fragment intensities are at least on par with existing tools. Additional sequence-based properties can also be predicted by AlphaPeptDeep, as demonstrated with a HLA peptide prediction model to improve HLA peptide identification for data-independent acquisition (https://github.com/MannLabs/PeptDeep-HLA).

- Röst et al. (2014), Demichev et al. (2019) — OpenSWATH and DIA-NN.

OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Röst et al. (2014). *Nature Biotechnology*. doi:10.1038/nbt.2841

DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Demichev et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0638-x

> We present an easy-to-use integrated software suite, DIA-NN, that exploits deep neural networks and new quantification and signal correction strategies for the processing of data-independent acquisition (DIA) proteomics experiments. DIA-NN improves the identification and quantification performance in conventional DIA proteomic applications, and is particularly beneficial for high-throughput applications, as it is fast and enables deep and confident proteome coverage when used in combination with fast chromatographic methods.

B.8.5 Statistics

- Keller et al. (2002) — PeptideProphet.

Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Keller et al. (2002). *Analytical Chemistry*. doi:10.1021/ac025747h

> We present a statistical model to estimate the accuracy of peptide assignments to tandem mass (MS/MS) spectra made by database search applications such as SEQUEST. Employing the expectation maximization algorithm, the analysis learns to distinguish correct from incorrect database search results, computing probabilities that peptide assignments to spectra are correct based upon database search scores and the number of tryptic termini of peptides. Using SEQUEST search results for spectra generated from a sample of known protein components, we demonstrate that the computed probabilities are accurate and have high power to discriminate between correctly and incorrectly assigned peptides. This analysis makes it possible to filter large volumes of MS/MS database search results with predictable false identification error rates and can serve as a common standard by which the results of different research groups are compared.

- Käll et al. (2007) — Percolator.

Semi-supervised learning for peptide identification from shotgun proteomics datasets. Käll et al. (2007). *Nature Methods*. doi:10.1038/nmeth1113

> Shotgun proteomics uses liquid chromatography-tandem mass spectrometry to identify proteins in complex biological samples. We describe an algorithm, called Percolator, for improving the rate of confident peptide identifications from a collection of tandem mass spectra. Percolator uses semi-supervised machine learning to discriminate between correct and decoy spectrum identifications, correctly assigning peptides to 17% more spectra from a tryptic Saccharomyces cerevisiae dataset, and up to 77% more spectra from non-tryptic digests, relative to a fully supervised approach.

- Nesvizhskii et al. (2003) — ProteinProphet and protein inference.

  A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. Nesvizhskii et al. (2003). *Analytical Chemistry*. doi:10.1021/ac0341261

  > A statistical model is presented for computing probabilities that proteins are present in a sample on the basis of peptides assigned to tandem mass (MS/MS) spectra acquired from a proteolytic digest of the sample. Peptides that correspond to more than a single protein in the sequence database are apportioned among all corresponding proteins, and a minimal protein list sufficient to account for the observed peptide assignments is derived using the expectation-maximization algorithm. Using peptide assignments to spectra generated from a sample of 18 purified proteins, as well as complex H. influenzae and Halobacterium samples, the model is shown to produce probabilities that are accurate and have high power to discriminate correct from incorrect protein identifications. This method allows filtering of large-scale proteomics data sets with predictable sensitivity and false positive identification error rates. Fast, consistent, and transparent, it provides a standard for publishing large-scale protein identification data sets in the literature and for comparing the results obtained from different experiments.

- Savitski et al. (2015) — protein-level FDR.

  A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Savitski et al. (2015). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.m114.046995

  > Calculating the number of confidently identified proteins and estimating false discovery rate (FDR) is a challenge when analyzing very large proteomic data sets such as entire human proteomes. Biological and technical heterogeneity in proteomic experiments further add to the challenge and there are strong differences in opinion regarding the conceptual validity of a protein FDR and no consensus regarding the methodology for protein FDR determination. There are also limitations inherent to the widely used classic target-decoy strategy that particularly show when analyzing very large data sets and that lead to a strong over-representation of decoy identifications. In this study, we investigated the merits of the classic, as well as a novel target-decoy-based protein FDR estimation approach, taking advantage of a heterogeneous data collection comprised of ~19,000 LC-MS/MS runs deposited in ProteomicsDB (https://www.proteomicsdb.org). The "picked" protein FDR approach treats target and decoy sequences of the same protein as a pair rather than as individual entities and chooses either the target or the decoy sequence depending on which receives the highest score. We investigated the performance of this approach in combination with q-value based peptide scoring to normalize sample-, instrument-, and search engine-specific differences. The "picked" target-decoy strategy performed best when protein scoring was based on the best peptide q-value for each protein yielding a stable number of true positive protein identifications over a wide range of q-value thresholds. We show that this simple and unbiased strategy eliminates a conceptual issue in the commonly used "classic" protein FDR approach that causes overprediction of false-positive protein identification in large data sets. The approach scales from small to very large data sets without losing performance, consistently increases the number of true-positive protein identifications and is readily implemented in proteomics analysis software.

- Choi et al. (2014) — MSstats.

  MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Choi et al. (2014). *Bioinformatics*. doi:10.1093/bioinformatics/btu305

  > Unlabelled MSstats is an R package for statistical relative quantification of proteins and peptides in mass spectrometry-based proteomics. Version 2.0 of MSstats supports label-free and label-based experimental workflows and data-dependent, targeted and data-independent spectral acquisition. It takes as input identified and quantified spectral peaks, and outputs a list of differentially abundant peptides or proteins, or summaries of peptide or protein relative abundance. MSstats relies on a flexible family of linear mixed models. Availability and implementation The code, the documentation and example datasets are available open-source at www.msstats.org under the Artistic-2.0 license. The package can be downloaded from www.msstats.org or from Bioconductor www.bioconductor.org and used in an R command line workflow. The package can also be accessed as an external tool in Skyline (Broudy et al., 2014) and used via graphical user interface.

B.8.6 Beyond abundance

- Olsen et al. (2006), Beausoleil et al. (2006) — phosphoproteomics and site localisation.

  Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. Olsen et al. (2006). *Cell*. doi:10.1016/j.cell.2006.09.026

  > Cell signaling mechanisms often transmit information via posttranslational protein modifications, most importantly reversible protein phosphorylation. Here we develop and apply a general mass spectrometric technology for identification and quantitation of phosphorylation sites as a function of stimulus, time, and subcellular location. We have detected 6,600 phosphorylation sites on 2,244 proteins and have determined their temporal dynamics after stimulating HeLa cells with epidermal growth factor (EGF) and recorded them in the Phosida database. Fourteen percent of phosphorylation sites are modulated at least 2-fold by EGF, and these were classified by their temporal profiles. Surprisingly, a majority of proteins contain multiple phosphorylation sites showing different kinetics, suggesting that they serve as platforms for integrating signals. In addition to protein kinase cascades, the targets of reversible phosphorylation include ubiquitin ligases, guanine nucleotide exchange factors, and at least 46 different transcriptional regulators. The dynamic phosphoproteome provides a missing link in a global, integrative view of cellular regulation.

  A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Beausoleil et al. (2006). *Nature Biotechnology*. doi:10.1038/nbt1240

  > Data analysis and interpretation remain major logistical challenges when attempting to identify large numbers of protein phosphorylation sites by nanoscale reverse-phase liquid chromatography/tandem mass spectrometry (LC-MS/MS) (Supplementary Figure 1 online). In this report we address challenges that are often only addressable by laborious manual validation, including data set error, data set sensitivity and phosphorylation site localization. We provide a large-scale phosphorylation data set with a measured error rate as determined by the target-decoy approach, we demonstrate an approach to maximize data set sensitivity by efficiently distracting incorrect peptide spectral matches (PSMs), and we present a probability-based score, the Ascore, that measures the probability of correct phosphorylation site localization based on the presence and intensity of site-determining ions in MS/MS spectra. We applied our methods in a fully automated fashion to nocodazole-arrested HeLa cell lysate where we identified 1,761 nonredundant phosphorylation sites from 491 proteins with a peptide false-positive rate of 1.3%.

- Gavin et al. (2002), Ho et al. (2002), Huttlin et al. (2021) — AP-MS maps of yeast and human complexes.

  Functional organization of the yeast proteome by systematic analysis of protein complexes. Gavin et al. (2002). *Nature*. doi:10.1038/415141a

  > Most cellular processes are carried out by multiprotein complexes. The identification and analysis of their components provides insight into how the ensemble of expressed proteins (proteome) is organized into functional units. We used tandem-affinity purification (TAP) and mass spectrometry in a large-scale approach to characterize multiprotein complexes in Saccharomyces cerevisiae. We processed 1,739 genes, including 1,143 human orthologues of relevance to human biology, and purified 589 protein assemblies. Bioinformatic analysis of these assemblies defined 232 distinct multiprotein complexes and proposed new cellular roles for 344 proteins, including 231 proteins with no previous functional annotation. Comparison of yeast and human complexes showed that conservation across species extends from single proteins to their molecular environment. Our analysis provides an outline of the eukaryotic proteome as a network of protein complexes at a level of organization beyond binary interactions. This higher-order map contains fundamental biological information and offers the context for a more reasoned and informed approach to drug discovery.

  Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Ho et al. (2002). *Nature*. doi:10.1038/415180a

  > The recent abundance of genome sequence data has brought an urgent need for systematic proteomics to decipher the encoded protein networks that dictate cellular function. To date, generation of large-scale protein-protein interaction maps has relied on the yeast two-hybrid system, which detects binary interactions through activation of reporter gene expression. With the advent of ultrasensitive mass spectrometric protein identification methods, it is feasible to identify directly protein complexes on

> a proteome-wide scale. Here we report, using the budding yeast Saccharomyces cerevisiae as a test case, an example of this approach, which we term high-throughput mass spectrometric protein complex identification (HMS-PCI). Beginning with 10% of predicted yeast proteins as baits, we detected 3,617 associated proteins covering 25% of the yeast proteome. Numerous protein complexes were identified, including many new interactions in various signalling pathways and in the DNA damage response. Comparison of the HMS-PCI data set with interactions reported in the literature revealed an average threefold higher success rate in detection of known complexes compared with large-scale two-hybrid studies. Given the high degree of connectivity observed in this study, even partial HMS-PCI coverage of complex proteomes, including that of humans, should allow comprehensive identification of cellular networks.

Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Huttlin et al. (2021). *Cell*. doi:10.1016/j.cell.2021.04.011

> Thousands of interactions assemble proteins into modules that impart spatial and functional organization to the cellular proteome. Through affinity-purification mass spectrometry, we have created two proteome-scale, cell-line-specific interaction networks. The first, BioPlex 3.0, results from affinity purification of 10,128 human proteins-half the proteome-in 293T cells and includes 118,162 interactions among 14,586 proteins. The second results from 5,522 immunoprecipitations in HCT116 cells. These networks model the interactome whose structure encodes protein function, localization, and complex membership. Comparison across cell lines validates thousands of interactions and reveals extensive customization. Whereas shared interactions reside in core complexes and involve essential proteins, cell-specific interactions link these complexes, "rewiring" subnetworks within each cell's interactome. Interactions covary among proteins of shared function as the proteome remodels to produce each cell's phenotype. Viewable interactively online through BioPlexExplorer, these networks define principles of proteome organization and enable unknown protein characterization.

- Choi et al. (2010), Mellacheruvu et al. (2013) — scoring AP-MS data.

SAINT: probabilistic scoring of affinity purification–mass spectrometry data. Choi et al. (2010). *Nature Methods*. doi:10.1038/nmeth.1541

> We present 'significance analysis of interactome' (SAINT), a computational tool that assigns confidence scores to protein-protein interaction data generated using affinity purification-mass spectrometry (AP-MS). The method uses label-free quantitative data and constructs separate distributions for true and false interactions to derive the probability of a bona fide protein-protein interaction. We show that SAINT is applicable to data of different scales and protein connectivity and allows transparent analysis of AP-MS data.

The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. Mellacheruvu et al. (2013). *Nature Methods*. doi:10.1038/nmeth.2557

> Affinity purification coupled with mass spectrometry (AP-MS) is a widely used approach for the identification of protein-protein interactions. However, for any given protein of interest, determining which of the identified polypeptides represent bona fide interactors versus those that are background contaminants (for example, proteins that interact with the solid-phase support, affinity reagent or epitope tag) is a challenging task. The standard approach is to identify nonspecific interactions using one or more negative-control purifications, but many small-scale AP-MS studies do not capture a complete, accurate background protein set when available controls are limited. Fortunately, negative controls are largely bait independent. Hence, aggregating negative controls from multiple AP-MS studies can increase coverage and improve the characterization of background associated with a given experimental protocol. Here we present the contaminant repository for affinity purification (the CRAPome) and describe its use for scoring protein-protein interactions. The repository (currently available for Homo sapiens and Saccharomyces cerevisiae) and computational tools are freely accessible at http://www.crapome.org/.

- Roux et al. (2012) — BioID proximity labelling.

A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. Roux et al. (2012). *Journal of Cell Biology*. doi:10.1083/jcb.201112098

> We have developed a new technique for proximity-dependent labeling of proteins in eukaryotic cells. Named BioID for proximity-dependent biotin identification, this approach is based on fusion of a promiscuous Escherichia coli biotin protein ligase to a targeting protein. BioID features proximity-dependent biotinylation of proteins that are near-neighbors of the fusion protein. Biotinylated proteins may be isolated by affinity capture and identified by mass spectrometry. We apply BioID to lamin-A (LaA), a well-characterized intermediate filament protein that is a constituent of the nuclear lamina, an important structural element of the nuclear envelope (NE). We identify multiple proteins that associate with and/or are proximate to LaA in vivo. The most abundant of these include known interactors of LaA that are localized to the NE, as well as a new NE-associated protein named SLAP75. Our results suggest BioID is a useful and generally applicable method to screen for both interacting and neighboring proteins in their native cellular environment.

- Havugimana et al. (2012), Wan et al. (2015) — co-fractionation MS.

A Census of Human Soluble Protein Complexes. Havugimana et al. (2012). *Cell*. doi:10.1016/j.cell.2012.08.011

> Cellular processes often depend on stable physical associations between proteins. Despite recent progress, knowledge of the composition of human protein complexes remains limited. To close this gap, we applied an integrative global proteomic profiling approach, based on chromatographic separation of cultured human cell extracts into more than one thousand biochemical fractions that were subsequently analyzed by quantitative tandem mass spectrometry, to systematically identify a network of 13,993 high-confidence physical interactions among 3,006 stably associated soluble human proteins. Most of the 622 putative protein complexes we report are linked to core biological processes and encompass both candidate disease genes and unannotated proteins to inform on mechanism. Strikingly, whereas larger multiprotein assemblies tend to be more extensively annotated and evolutionarily conserved, human protein complexes with five or fewer subunits are far more likely to be functionally unannotated or restricted to vertebrates, suggesting more recent functional innovations.

Panorama of ancient metazoan macromolecular complexes. Wan et al. (2015). *Nature*. doi:10.1038/nature14877

> Macromolecular complexes are essential to conserved biological processes, but their prevalence across animals is unclear. By combining extensive biochemical fractionation with quantitative mass spectrometry, here we directly examined the composition of soluble multiprotein complexes among diverse metazoan models. Using an integrative approach, we generated a draft conservation map consisting of more than one million putative high-confidence co-complex interactions for species with fully sequenced genomes that encompasses functional modules present broadly across all extant animals. Clustering reveals a spectrum of conservation, ranging from ancient eukaryotic assemblies that have probably served cellular housekeeping roles for at least one billion years, ancestral complexes that have accrued contemporary components, and rarer metazoan innovations linked to multicellularity. We validated these projections by independent co-fractionation experiments in evolutionarily distant species, affinity purification and functional analyses. The comprehensiveness, centrality and modularity of these reconstructed interactomes reflect their fundamental mechanistic importance and adaptive value to animal cell systems.

- Drew et al. (2021), Fischer et al. (2025) — hu.MAP 2.0 and 3.0.

hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Drew et al. (2021). *Molecular Systems Biology*. doi:10.15252/msb.202010016

> A general principle of biology is the self-assembly of proteins into functional complexes. Characterizing their composition is, therefore, required for our understanding of cellular functions. Unfortunately, we lack knowledge of the comprehensive set of identities of protein complexes in human cells. To address this gap, we developed a machine learning framework to identify protein complexes in over 15,000 mass spectrometry experiments which resulted in the identification of nearly 7,000 physical assemblies. We show our resource, hu.MAP 2.0, is more accurate and comprehensive than previous state of the art high-throughput protein complex resources and gives rise to many new hypotheses, including for 274 completely uncharacterized proteins. Further, we identify 253 promiscuous proteins that participate in multiple complexes pointing to possible moonlighting roles. We have made hu.MAP 2.0 easily searchable in a web interface (http://humap2.proteincomplexes.org/), which will be a valuable resource for researchers across a broad range of interests including systems biology, structural biology, and molecular explanations of disease.

hu.MAP3.0: atlas of human protein complexes by integration of >25,000 proteomic experiments. Fischer et al. (2025). *Molecular Systems Biology*. doi:10.1038/s44320-025-00121-5

> Macromolecular protein complexes carry out most cellular functions. Unfortunately, we lack the subunit composition for many human protein complexes. To address this gap we integrated >25,000 mass spectrometry experiments using a machine learning approach to identify >15,000 human protein complexes. We show our map of protein complexes is highly accurate and more comprehensive than previous maps, placing nearly 70% of human proteins into their physical contexts. We globally characterize our complexes using mass spectrometry based protein covariation data (ProteomeHD.2) and identify covarying complexes suggesting common functional associations. hu.MAP3.0 generates testable functional hypotheses for 472 uncharacterized proteins which we support using AlphaFold modeling. Additionally, we use AlphaFold modeling to identify 5871 mutually exclusive proteins in hu.MAP3.0 complexes suggesting complexes serve different functional roles depending on their subunit composition. We identify expression as the primary way cells and organisms relieve the conflict of mutually exclusive subunits. Finally, we import our complexes to EMBL-EBI's Complex Portal (https://www.ebi.ac.uk/complexportal/home) and provide complexes through our hu.MAP3.0 web interface (https://humap3.proteincomplexes.org/). We expect our resource to be highly impactful to the broader research community.

- O'Reilly and Rappsilber (2018) — cross-linking MS.

Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. O'Reilly & Rappsilber (2018). *Nature Structural & Molecular Biology*. doi:10.1038/s41594-018-0147-0

> Over the past decade, cross-linking mass spectrometry (CLMS) has developed into a robust and flexible tool that provides medium-resolution structural information. CLMS data provide a measure of the proximity of amino acid residues and thus offer information on the folds of proteins and the topology of their complexes. Here, we highlight notable successes of this technique as well as common pipelines. Novel CLMS applications, such as in-cell cross-linking, probing conformational changes and tertiary-structure determination, are now beginning to make contributions to molecular biology and the emerging fields of structural systems biology and interactomics.

- Schopper et al. (2017), Savitski et al. (2014) — LiP-MS and thermal proteome profiling.

Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. Schopper et al. (2017). *Nature Protocols*. doi:10.1038/nprot.2017.100

> Protein structural changes induced by external perturbations or internal cues can profoundly influence protein activity and thus modulate cellular physiology. A number of biophysical approaches are available to probe protein structural changes, but these are not applicable to a whole proteome in a biological extract. Limited proteolysis-coupled mass spectrometry (LiP-MS) is a recently developed proteomics approach that enables the identification of protein structural changes directly in their complex biological context on a proteome-wide scale. After perturbations of interest, proteome extracts are subjected to a double-protease digestion step with a nonspecific protease applied under native conditions, followed by complete digestion with the sequence-specific protease trypsin under denaturing conditions. This sequential treatment generates structure-specific peptides amenable to bottom-up MS analysis. Next, a proteomics workflow involving shotgun or targeted MS and label-free quantification is applied to measure structure-dependent proteolytic patterns directly in the proteome extract. Possible applications of LiP-MS include discovery of perturbation-induced protein structural alterations, identification of drug targets, detection of disease-associated protein structural states, and analysis of protein aggregates directly in biological samples. The approach also enables identification of the specific protein regions involved in the structural transition or affected by the binding event. Sample preparation takes approximately 2 d, followed by one to several days of MS and data analysis time, depending on the number of samples analyzed. Scientists with basic biochemistry training can implement the sample preparation steps. MS measurement and data analysis require a background in proteomics.

Tracking cancer drugs in living cells by thermal profiling of the proteome. Savitski et al. (2014). *Science*. doi:10.1126/science.1255784

> The thermal stability of proteins can be used to assess ligand binding in living cells. We have generalized this concept by determining the thermal profiles of more than 7000 proteins in human cells by means of mass spectrometry. Monitoring the effects of small-molecule ligands on the profiles delineated more than 50 targets for the kinase inhibitor staurosporine. We identified the heme biosynthesis enzyme

ferrochelatase as a target of kinase inhibitors and suggest that its inhibition causes the phototoxicity observed with vemurafenib and alectinib. Thermal shifts were also observed for downstream effectors of drug treatment. In live cells, dasatinib induced shifts in BCR-ABL pathway proteins, including CRK/CRKL. Thermal proteome profiling provides an unbiased measure of drug-target engagement and facilitates identification of markers for drug efficacy and toxicity.

- Christoforou et al. (2016), Thul et al. (2017) — spatial proteomics of organelles and cells.

  A draft map of the mouse pluripotent stem cell spatial proteome. Christoforou et al. (2016). *Nature Communications*. doi:10.1038/ncomms9992

  > Knowledge of the subcellular distribution of proteins is vital for understanding cellular mechanisms. Capturing the subcellular proteome in a single experiment has proven challenging, with studies focusing on specific compartments or assigning proteins to subcellular niches with low resolution and/or accuracy. Here we introduce hyperLOPIT, a method that couples extensive fractionation, quantitative high-resolution accurate mass spectrometry with multivariate data analysis. We apply hyperLOPIT to a pluripotent stem cell population whose subcellular proteome has not been extensively studied. We provide localization data on over 5,000 proteins with unprecedented spatial resolution to reveal the organization of organelles, sub-organellar compartments, protein complexes, functional networks and steady-state dynamics of proteins and unexpected subcellular locations. The method paves the way for characterizing the impact of post-transcriptional and post-translational modification on protein location and studies involving proteome-level locational changes on cellular perturbation. An interactive open-source resource is presented that enables exploration of these data.

  A subcellular map of the human proteome. Thul et al. (2017). *Science*. doi:10.1126/science.aal3321

  > Resolving the spatial distribution of the human proteome at a subcellular level can greatly increase our understanding of human biology and disease. Here we present a comprehensive image-based map of subcellular protein distribution, the Cell Atlas, built by integrating transcriptomics and antibody-based immunofluorescence microscopy with validation by mass spectrometry. Mapping the in situ localization of 12,003 human proteins at a single-cell level to 30 subcellular structures enabled the definition of the proteomes of 13 major organelles. Exploration of the proteomes revealed single-cell variations in abundance or spatial distribution and localization of about half of the proteins to multiple compartments. This subcellular map can be used to refine existing protein-protein interaction networks and provides an important resource to deconvolute the highly complex architecture of the human cell.

- Caprioli et al. (1997), Mund et al. (2022), Budnik et al. (2018) — MS imaging, deep visual proteomics and single-cell proteomics.

  Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. Caprioli et al. (1997). *Analytical Chemistry*. doi:10.1021/ac970888i

  > Matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS) has been used to generate ion images of samples in one or more mass-to-charge (m/z) values, providing the capability of mapping specific molecules to two-dimensional coordinates of the original sample. The high sensitivity of the technique (low-femtomole to attomole levels for proteins and peptides) allows the study of organized biochemical processes occurring in, for example, mammalian tissue sections. The mass spectrometer is used to determine the molecular weights of the molecular in the surface layers of the tissue. Molecules desorbed from the sample typically are singly protonated, giving an ion at (M + H)+, where M is the molecular mass. The procedure involves coating the tissue section, or a blotted imprint of the section, with a thin layer of energy-absorbing matrix and then analyzing the sample to produce an ordered array of mass spectra, each containing nominal m/z values typically covering a range of over 50,000 Da. Images can be displayed in individual m/z values as a selected ion image, which would localize individual compounds in the tissue, or as summed ion images. MALDI ion images of tissue sections can be obtained directly from tissue slices following preparative steps, and this is demonstrated for the mapping of insulin contained in an islet in a section of rat pancreas, hormone peptides in a small area of a section of rat pituitary, and a small protein bound to the membrane of human mucosa cells. Alternatively, imprints of the tissue can be analyzed by blotting the tissue sections on specially prepared targets containing an adsorbent material, e.g., C-18 coated resin beads. Peptides and small proteins bind to the C-18 and create a positive imprint of the tissue which can then be imaged by the mass spectrometer. This is demonstrated for the MALDI ion image analysis of regions of rat splenic pancreas and for an area of rat pituitary traversing the anterior, intermediate, and posterior regions where localized peptides were mapped. In a single spectrum from the

anterior/intermediate lobe of a rat pituitary print, over 50 ions corresponding to the peptides present in this tissue were observed as well as precursors, isoforms, and metabolic fragments.

Deep Visual Proteomics defines single-cell identity and heterogeneity. Mund et al. (2022). *Nature Biotechnology*. doi:10.1038/s41587-022-01302-5

> Despite the availabilty of imaging-based and mass-spectrometry-based methods for spatial proteomics, a key challenge remains connecting images with single-cell-resolution protein abundance measurements. Here, we introduce Deep Visual Proteomics (DVP), which combines artificial-intelligence-driven image analysis of cellular phenotypes with automated single-cell or single-nucleus laser microdissection and ultra-high-sensitivity mass spectrometry. DVP links protein abundance to complex cellular or subcellular phenotypes while preserving spatial context. By individually excising nuclei from cell culture, we classified distinct cell states with proteomic profiles defined by known and uncharacterized proteins. In an archived primary melanoma tissue, DVP identified spatially resolved proteome changes as normal melanocytes transition to fully invasive melanoma, revealing pathways that change in a spatial manner as cancer progresses, such as mRNA splicing dysregulation in metastatic vertical growth that coincides with reduced interferon signaling and antigen presentation. The ability of DVP to retain precise spatial proteomic information in the tissue context has implications for the molecular profiling of clinical samples.

SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Budnik et al. (2018). *Genome Biology*. doi:10.1186/s13059-018-1547-5

> Some exciting biological questions require quantifying thousands of proteins in single cells. To achieve this goal, we develop Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS) and validate its ability to identify distinct human cancer cell types based on their proteomes. We use SCoPE-MS to quantify over a thousand proteins in differentiating mouse embryonic stem cells. The single-cell proteomes enable us to deconstruct cell populations and infer protein abundance relationships. Comparison between single-cell proteomes and transcriptomes indicates coordinated mRNA and protein covariation, yet many genes exhibit functionally concerted and distinct regulatory patterns at the mRNA and the protein level.

B.8.7 Non-MS proteomics and resources

- Thul et al. (2017), Uhlen et al. (2017), Karlsson et al. (2021) — the Human Protein Atlas.

A subcellular map of the human proteome. Thul et al. (2017). *Science*. doi:10.1126/science.aal3321

> Resolving the spatial distribution of the human proteome at a subcellular level can greatly increase our understanding of human biology and disease. Here we present a comprehensive image-based map of subcellular protein distribution, the Cell Atlas, built by integrating transcriptomics and antibody-based immunofluorescence microscopy with validation by mass spectrometry. Mapping the in situ localization of 12,003 human proteins at a single-cell level to 30 subcellular structures enabled the definition of the proteomes of 13 major organelles. Exploration of the proteomes revealed single-cell variations in abundance or spatial distribution and localization of about half of the proteins to multiple compartments. This subcellular map can be used to refine existing protein-protein interaction networks and provides an important resource to deconvolute the highly complex architecture of the human cell.

A pathology atlas of the human cancer transcriptome. Uhlen et al. (2017). *Science*. doi:10.1126/science.aan2507

> Cancer is one of the leading causes of death, and there is great interest in understanding the underlying molecular mechanisms involved in the pathogenesis and progression of individual tumors. We used systems-level approaches to analyze the genome-wide transcriptome of the protein-coding genes of 17 major cancer types with respect to clinical outcome. A general pattern emerged: Shorter patient survival was associated with up-regulation of genes involved in cell growth and with down-regulation of genes involved in cellular differentiation. Using genome-scale metabolic models, we show that cancer patients have widespread metabolic heterogeneity, highlighting the need for precise and personalized medicine for cancer treatment. All data are presented in an interactive open-access database (www.proteinatlas.org/pathology) to allow genome-wide exploration of the impact of individual proteins on clinical outcomes.

A single–cell type transcriptomics map of human tissues. Karlsson et al. (2021). *Science Advances*. doi:10.1126/sciadv.abh2169

> Advances in molecular profiling have opened up the possibility to map the expression of genes in cells, tissues, and organs in the human body. Here, we combined single-cell transcriptomics analysis with spatial antibody-based protein profiling to create a high-resolution single-cell type map of human tissues. An open access atlas has been launched to allow researchers to explore the expression of human protein-coding genes in 192 individual cell type clusters. An expression specificity classification was performed to determine the number of genes elevated in each cell type, allowing comparisons with bulk transcriptomics data. The analysis highlights distinct expression clusters corresponding to cell types sharing similar functions, both within the same organs and between organs.

- Uhlen et al. (2016) — validation of antibodies.

  A proposal for validation of antibodies. Uhlen et al. (2016). *Nature Methods*. doi:10.1038/nmeth.3995

- Assarsson et al. (2014), Gold et al. (2010) — PEA and SomaScan.

  Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. Assarsson et al. (2014). *PLoS ONE*. doi:10.1371/journal.pone.0095192

  > Medical research is developing an ever greater need for comprehensive high-quality data generation to realize the promises of personalized health care based on molecular biomarkers. The nucleic acid proximity-based methods proximity ligation and proximity extension assays have, with their dual reporters, shown potential to relieve the shortcomings of antibodies and their inherent cross-reactivity in multiplex protein quantification applications. The aim of the present study was to develop a robust 96-plex immunoassay based on the proximity extension assay (PEA) for improved high throughput detection of protein biomarkers. This was enabled by: (1) a modified design leading to a reduced number of pipetting steps compared to the existing PEA protocol, as well as improved intra-assay precision; (2) a new enzymatic system that uses a hyper-thermostabile enzyme, Pwo, for uniting the two probes allowing for room temperature addition of all reagents and improved the sensitivity; (3) introduction of an inter-plate control and a new normalization procedure leading to improved inter-assay precision (reproducibility). The multiplex proximity extension assay was found to perform well in complex samples, such as serum and plasma, and also in xenografted mice and resuspended dried blood spots, consuming only 1 μL sample per test. All-in-all, the development of the current multiplex technique is a step toward robust high throughput protein marker discovery and research.

  Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. Gold et al. (2010). *PLoS ONE*. doi:10.1371/journal.pone.0015004

  > Background The interrogation of proteomes ("proteomics") in a highly multiplexed and efficient manner remains a coveted and challenging goal in biology and medicine. Methodology/principal findings We present a new aptamer-based proteomic technology for biomarker discovery capable of simultaneously measuring thousands of proteins from small sample volumes (15 μL of serum or plasma). Our current assay measures 813 proteins with low limits of detection (1 pM median), 7 logs of overall dynamic range (~100 fM-1 μM), and 5% median coefficient of variation. This technology is enabled by a new generation of aptamers that contain chemically modified nucleotides, which greatly expand the physicochemical diversity of the large randomized nucleic acid libraries from which the aptamers are selected. Proteins in complex matrices such as plasma are measured with a process that transforms a signature of protein concentrations into a corresponding signature of DNA aptamer concentrations, which is quantified on a DNA microarray. Our assay takes advantage of the dual nature of aptamers as both folded protein-binding entities with defined shapes and unique nucleotide sequences recognizable by specific hybridization probes. To demonstrate the utility of our proteomics biomarker discovery technology, we applied it to a clinical study of chronic kidney disease (CKD). We identified two well known CKD biomarkers as well as an additional 58 potential CKD biomarkers. These results demonstrate the potential utility of our technology to rapidly discover unique protein signatures characteristic of various disease states. Conclusions/significance We describe a versatile and powerful tool that allows large-scale comparison of proteome profiles among discrete populations. This unbiased and highly multiplexed search engine will enable the discovery of novel biomarkers in a manner that is unencumbered by our incomplete knowledge of biology, thereby helping to advance the next generation of evidence-based medicine.

- Sun et al. (2023) — plasma proteomics in the UK Biobank.

  Plasma proteomic associations with genetics and health in the UK Biobank. Sun et al. (2023). *Nature*. doi:10.1038/s41586-023-06592-6

  > The Pharma Proteomics Project is a precompetitive biopharmaceutical consortium characterizing the plasma proteomic profiles of 54,219 UK Biobank participants. Here we provide a detailed summary of this initiative, including technical and biological validations, insights into proteomic disease signatures, and prediction modelling for various demographic and health indicators. We present comprehensive protein quantitative trait locus (pQTL) mapping of 2,923 proteins that identifies 14,287 primary genetic associations, of which 81% are previously undescribed, alongside ancestry-specific pQTL mapping in non-European individuals. The study provides an updated characterization of the genetic architecture of the plasma proteome, contextualized with projected pQTL discovery rates as sample sizes and proteomic assay coverages increase over time. We offer extensive insights into trans pQTLs across multiple biological domains, highlight genetic influences on ligand-receptor interactions and pathway perturbations across a diverse collection of cytokines and complement networks, and illustrate long-range epistatic effects of ABO blood group and FUT2 secretor status on proteins with gastrointestinal tissue-enriched expression. We demonstrate the utility of these data for drug discovery by extending the genetic proxied effects of protein targets, such as PCSK9, on additional endpoints, and disentangle specific genes and proteins perturbed at loci associated with COVID-19 susceptibility. This public-private partnership provides the scientific community with an open-access proteomics resource of considerable breadth and depth to help to elucidate the biological mechanisms underlying proteo-genomic discoveries and accelerate the development of biomarkers, predictive models and therapeutics 1 .

- Goltsev et al. (2018), Giesen et al. (2014) — CODEX and imaging mass cytometry.

  Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Goltsev et al. (2018). *Cell*. doi:10.1016/j.cell.2018.07.010

  > A highly multiplexed cytometric imaging approach, termed co-detection by indexing (CODEX), is used here to create multiplexed datasets of normal and lupus (MRL/lpr) murine spleens. CODEX iteratively visualizes antibody binding events using DNA barcodes, fluorescent dNTP analogs, and an in situ polymerization-based indexing procedure. An algorithmic pipeline for single-cell antigen quantification in tightly packed tissues was developed and used to overlay well-known morphological features with de novo characterization of lymphoid tissue architecture at a single-cell and cellular neighborhood levels. We observed an unexpected, profound impact of the cellular neighborhood on the expression of protein receptors on immune cells. By comparing normal murine spleen to spleens from animals with systemic autoimmune disease (MRL/lpr), extensive and previously uncharacterized splenic cell-interaction dynamics in the healthy versus diseased state was observed. The fidelity of multiplexed spatial cytometry demonstrated here allows for quantitative systemic characterization of tissue architecture in normal and clinically aberrant samples.

  Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Giesen et al. (2014). *Nature Methods*. doi:10.1038/nmeth.2869

  > Mass cytometry enables high-dimensional, single-cell analysis of cell type and state. In mass cytometry, rare earth metals are used as reporters on antibodies. Analysis of metal abundances using the mass cytometer allows determination of marker expression in individual cells. Mass cytometry has previously been applied only to cell suspensions. To gain spatial information, we have coupled immunohistochemical and immunocytochemical methods with high-resolution laser ablation to CyTOF mass cytometry. This approach enables the simultaneous imaging of 32 proteins and protein modifications at subcellular resolution; with the availability of additional isotopes, measurement of over 100 markers will be possible. We applied imaging mass cytometry to human breast cancer samples, allowing delineation of cell subpopulations and cell-cell interactions and highlighting tumor heterogeneity. Imaging mass cytometry complements existing imaging approaches. It will enable basic studies of tissue heterogeneity and function and support the transition of medicine toward individualized molecularly targeted diagnosis and therapies.

- Alfaro et al. (2021) — single-molecule protein sequencing.

  The emerging landscape of single-molecule protein sequencing technologies. Alfaro et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01143-1

> Single-cell profiling methods have had a profound impact on the understanding of cellular heterogeneity. While genomes and transcriptomes can be explored at the single-cell level, single-cell profiling of proteomes is not yet established. Here we describe new single-molecule protein sequencing and identification technologies alongside innovations in mass spectrometry that will eventually enable broad sequence coverage in single-cell profiling. These technologies will in turn facilitate biological discovery and open new avenues for ultrasensitive disease diagnostics.

- Perez-Riverol et al. (2024), Vizcaíno et al. (2014), Deutsch et al. (2022), Desiere (2006) — PRIDE, ProteomeXchange and PeptideAtlas.

The PRIDE database at 20 years: 2025 update. Perez-Riverol et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1011

> The PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) is the world's leading mass spectrometry (MS)-based proteomics data repository and one of the founding members of the ProteomeXchange consortium. This manuscript summarizes the developments in PRIDE resources and related tools for the last three years. The number of submitted datasets to PRIDE Archive (the archival component of PRIDE) has reached on average around 534 datasets per month. This has been possible thanks to continuous improvements in infrastructure such as a new file transfer protocol for very large datasets (Globus), a new data resubmission pipeline and an automatic dataset validation process. Additionally, we will highlight novel activities such as the availability of the PRIDE chatbot (based on the use of open-source Large Language Models), and our work to improve support for MS crosslinking datasets. Furthermore, we will describe how we have increased our efforts to reuse, reanalyze and disseminate high-quality proteomics data into added-value resources such as UniProt, Ensembl and Expression Atlas.

ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Vizcaíno et al. (2014). *Nature Biotechnology*. doi:10.1038/nbt.2839

The ProteomeXchange consortium at 10 years: 2023 update. Deutsch et al. (2022). *Nucleic Acids Research*. doi:10.1093/nar/gkac1040

> Mass spectrometry (MS) is by far the most used experimental approach in high-throughput proteomics. The ProteomeXchange (PX) consortium of proteomics resources (http://www.proteomexchange.org) was originally set up to standardize data submission and dissemination of public MS proteomics data. It is now 10 years since the initial data workflow was implemented. In this manuscript, we describe the main developments in PX since the previous update manuscript in Nucleic Acids Research was published in 2020. The six members of the Consortium are PRIDE, PeptideAtlas (including PASSEL), MassIVE, jPOST, iProX and Panorama Public. We report the current data submission statistics, showcasing that the number of datasets submitted to PX resources has continued to increase every year. As of June 2022, more than 34 233 datasets had been submitted to PX resources, and from those, 20 062 (58.6%) just in the last three years. We also report the development of the Universal Spectrum Identifiers and the improvements in capturing the experimental metadata annotations. In parallel, we highlight that data re-use activities of public datasets continue to increase, enabling connections between PX resources and other popular bioinformatics resources, novel research and also new data resources. Finally, we summarise the current state-of-the-art in data management practices for sensitive human (clinical) proteomics data.

The PeptideAtlas project. Desiere (2006). *Nucleic Acids Research*. doi:10.1093/nar/gkj040

> The completion of the sequencing of the human genome and the concurrent, rapid development of high-throughput proteomic methods have resulted in an increasing need for automated approaches to archive proteomic data in a repository that enables the exchange of data among researchers and also accurate integration with genomic data. PeptideAtlas (http://www.peptideatlas.org/) addresses these needs by identifying peptides by tandem mass spectrometry (MS/MS), statistically validating those identifications and then mapping identified sequences to the genomes of eukaryotic organisms. A meaningful comparison of data across different experiments generated by different groups using different types of instruments is enabled by the implementation of a uniform analytic process. This uniform statistical validation ensures a consistent and high-quality set of peptide and protein identifications. The raw data from many diverse proteomic experiments are made available in the associated PeptideAtlas repository in several formats. Here we present a summary of our process and details about the Human, Drosophila and Yeast PeptideAtlas builds.

- Martens et al. (2011), Taylor et al. (2007) — mzML and MIAPE.

  mzML—a Community Standard for Mass Spectrometry Data. Martens et al. (2011). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.r110.000133

  > Mass spectrometry is a fundamental tool for discovery and analysis in the life sciences. With the rapid advances in mass spectrometry technology and methods, it has become imperative to provide a standard output format for mass spectrometry data that will facilitate data sharing and analysis. Initially, the efforts to develop a standard format for mass spectrometry data resulted in multiple formats, each designed with a different underlying philosophy. To resolve the issues associated with having multiple formats, vendors, researchers, and software developers convened under the banner of the HUPO PSI to develop a single standard. The new data format incorporated many of the desirable technical attributes from the previous data formats, while adding a number of improvements, including features such as a controlled vocabulary with validation tools to ensure consistent usage of the format, improved support for selected reaction monitoring data, and immediately available implementations to facilitate rapid adoption by the community. The resulting standard data format, mzML, is a well tested open-source format for mass spectrometer output files that can be readily utilized by the community and easily adapted for incremental advances in mass spectrometry technology.

  The minimum information about a proteomics experiment (MIAPE). Taylor et al. (2007). *Nature Biotechnology*. doi:10.1038/nbt1329

  > Both the generation and the analysis of proteomics data are now widespread, and high-throughput approaches are commonplace. Protocols continue to increase in complexity as methods and technologies evolve and diversify. To encourage the standardized collection, integration, storage and dissemination of proteomics data, the Human Proteome Organization's Proteomics Standards Initiative develops guidance modules for reporting the use of techniques such as gel electrophoresis and mass spectrometry. This paper describes the processes and principles underpinning the development of these modules; discusses the ramifications for various interest groups such as experimentalists, funders, publishers and the private sector; addresses the issue of overlap with other reporting guidelines; and highlights the criticality of appropriate tools and resources in enabling 'MIAPE-compliant' reporting.

- Kim et al. (2014), Wilhelm et al. (2014), Ezkurdia, Vázquez, et al. (2014) — draft maps of the human proteome and their reanalysis.

  A draft map of the human proteome. Kim et al. (2014). *Nature*. doi:10.1038/nature13302

  > The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic analysis enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non-coding RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive web-based resource at http://www.humanproteomemap.org) will complement available human genome and transcriptome data to accelerate biomedical research in health and disease.

  Mass-spectrometry-based draft of the human proteome. Wilhelm et al. (2014). *Nature*. doi:10.1038/nature13319

  > Proteomes are characterized by large protein-abundance differences, cell-type- and time-dependent expression patterns and post-translational modifications, all of which carry biological information that is not accessible by genomics or transcriptomics. Here we present a mass-spectrometry-based draft of the human proteome and a public, high-performance, in-memory database for real-time analysis of terabytes of big data, called ProteomicsDB. The information assembled from human tissues, cell lines and body fluids enabled estimation of the size of the protein-coding genome, and identified organ-specific proteins and a large number of translated lincRNAs (long intergenic non-coding RNAs). Analysis of messenger RNA and protein-expression profiles of human tissues revealed conserved control of protein abundance, and integration of drug-sensitivity data enabled the identification of proteins predicting resistance or sensitivity. The proteome profiles also hold considerable promise for analysing the composition and

stoichiometry of protein complexes. ProteomicsDB thus enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

Analyzing the First Drafts of the Human Proteome. Ezkurdia et al. (2014). *Journal of Proteome Research*. doi:10.1021/pr500572z

> This letter analyzes two large-scale proteomics studies published in the same issue of Nature. At the time of the release, both studies were portrayed as draft maps of the human proteome and great advances in the field. As with the initial publication of the human genome, these papers have broad appeal and will no doubt lead to a great deal of further analysis by the scientific community. However, we were intrigued by the number of protein-coding genes detected by the two studies, numbers that far exceeded what has been reported for the multinational Human Proteome Project effort. We carried out a simple quality test on the data using the olfactory receptor family. A high-quality proteomics experiment that does not specifically analyze nasal tissues should not expect to detect many peptides for olfactory receptors. Neither of the studies carried out experiments on nasal tissues, yet we found peptide evidence for more than 100 olfactory receptors in the two studies. These results suggest that the two studies are substantially overestimating the number of protein coding genes they identify. We conclude that the experimental data from these two studies should be used with caution.

- Legrain et al. (2011), Omenn et al. (2024) — the Human Proteome Project.

The Human Proteome Project: Current State and Future Direction. Legrain et al. (2011). *Molecular & Cellular Proteomics*. doi:10.1074/mcp.m111.009993

> After the successful completion of the Human Genome Project, the Human Proteome Organization has recently officially launched a global Human Proteome Project (HPP), which is designed to map the entire human protein set. Given the lack of protein-level evidence for about 30% of the estimated 20,300 protein-coding genes, a systematic global effort will be necessary to achieve this goal with respect to protein abundance, distribution, subcellular localization, interaction with other biomolecules, and functions at specific time points. As a general experimental strategy, HPP research groups will use the three working pillars for HPP: mass spectrometry, antibody capture, and bioinformatics tools and knowledge bases. The HPP participants will take advantage of the output and cross-analyses from the ongoing Human Proteome Organization initiatives and a chromosome-centric protein mapping strategy, termed C-HPP, with which many national teams are currently engaged. In addition, numerous biologically driven and disease-oriented projects will be stimulated and facilitated by the HPP. Timely planning with proper governance of HPP will deliver a protein parts list, reagents, and tools for protein studies and analyses, and a stronger basis for personalized medicine. The Human Proteome Organization urges each national research funding agency and the scientific community at large to identify their preferred pathways to participate in aspects of this highly promising project in a HPP consortium of funders and investigators.

The 2023 Report on the Proteome from the HUPO Human Proteome Project. Omenn et al. (2024). *Journal of Proteome Research*. doi:10.1021/acs.jproteome.3c00591

> Since 2010, the Human Proteome Project (HPP), the flagship initiative of the Human Proteome Organization (HUPO), has pursued two goals: (1) to credibly identify the protein parts list and (2) to make proteomics an integral part of multiomics studies of human health and disease. The HPP relies on international collaboration, data sharing, standardized reanalysis of MS data sets by PeptideAtlas and MassIVE-KB using HPP Guidelines for quality assurance, integration and curation of MS and non-MS protein data by neXtProt, plus extensive use of antibody profiling carried out by the Human Protein Atlas. According to the neXtProt release 2023-04-18, protein expression has now been credibly detected (PE1) for 18,397 of the 19,778 neXtProt predicted proteins coded in the human genome (93%). Of these PE1 proteins, 17,453 were detected with mass spectrometry (MS) in accordance with HPP Guidelines and 944 by a variety of non-MS methods. The number of neXtProt PE2, PE3, and PE4 missing proteins now stands at 1381. Achieving the unambiguous identification of 93% of predicted proteins encoded from across all chromosomes represents remarkable experimental progress on the Human Proteome parts list. Meanwhile, there are several categories of predicted proteins that have proved resistant to detection regardless of protein-based methods used. Additionally there are some PE1-4 proteins that probably should be reclassified to PE5, specifically 21 LINC entries and ~30 HERV entries; these are being addressed in the present year. Applying proteomics in a wide array of biological and clinical studies ensures integration with other omics platforms as reported by the Biology and Disease-driven HPP teams and the antibody and pathology resource pillars. Current progress has positioned the HPP to transition to its Grand Challenge Project focused on determining the primary function(s) of every protein itself and in networks and pathways within the context of human health and disease.

- Ideker et al. (2001) — integrated genomic and proteomic analysis of a perturbed network.

  Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Ideker et al. (2001). *Science*. doi:10.1126/science.292.5518.929

  > We demonstrate an integrated approach to build, test, and refine a model of a cellular pathway, in which perturbations to critical pathway components are analyzed using DNA microarrays, quantitative proteomics, and databases of known physical interactions. Using this approach, we identify 997 messenger RNAs responding to 20 systematic perturbations of the yeast galactose-utilization pathway, provide evidence that approximately 15 of 289 detected proteins are regulated posttranscriptionally, and identify explicit physical interactions governing the cellular response to each perturbation. We refine the model through further iterations of perturbation and global measurements, suggesting hypotheses about the regulation of galactose utilization and physical interactions between this and a variety of other metabolic pathways.

B.9 Protein structure

Further papers for Chapter 9.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.9.1 CASP and evaluation

- Moult et al. (1995) — the first CASP.

  A large-scale experiment to assess protein structure prediction methods. Moult et al. (1995). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.340230303

- Kryshtafovych et al. (2021), Kryshtafovych et al. (2023), Yuan et al. (2025) — CASP14, CASP15 and CASP16.

  Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Kryshtafovych et al. (2021). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.26237

  > Critical assessment of structure prediction (CASP) is a community experiment to advance methods of computing three-dimensional protein structure from amino acid sequence. Core components are rigorous blind testing of methods and evaluation of the results by independent assessors. In the most recent experiment (CASP14), deep-learning methods from one research group consistently delivered computed structures rivaling the corresponding experimental ones in accuracy. In this sense, the results represent a solution to the classical protein-folding problem, at least for single proteins. The models have already been shown to be capable of providing solutions for problematic crystal structures, and there are broad implications for the rest of structural biology. Other research groups also substantially improved performance. Here, we describe these results and outline some of the many implications. Other related areas of CASP, including modeling of protein complexes, structure refinement, estimation of model accuracy, and prediction of inter-residue contacts and distances, are also described.

  Critical assessment of methods of protein structure prediction (CASP)—Round XV. Kryshtafovych et al. (2023). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.26617

  > Computing protein structure from amino acid sequence information has been a long-standing grand challenge. Critical assessment of structure prediction (CASP) conducts community experiments aimed at advancing solutions to this and related problems. Experiments are conducted every 2 years. The 2020 experiment (CASP14) saw major progress, with the second generation of deep learning methods delivering accuracy comparable with experiment for many single proteins. There is an expectation that these methods will have much wider application in computational structural biology. Here we summarize

results from the most recent experiment, CASP15, in 2022, with an emphasis on new deep learning-driven progress. Other papers in this special issue of proteins provide more detailed analysis. For single protein structures, the AlphaFold2 deep learning method is still superior to other approaches, but there are two points of note. First, although AlphaFold2 was the core of all the most successful methods, there was a wide variety of implementation and combination with other methods. Second, using the standard AlphaFold2 protocol and default parameters only produces the highest quality result for about two thirds of the targets, and more extensive sampling is required for the others. The major advance in this CASP is the enormous increase in the accuracy of computed protein complexes, achieved by the use of deep learning methods, although overall these do not fully match the performance for single proteins. Here too, AlphaFold2 based method perform best, and again more extensive sampling than the defaults is often required. Also of note are the encouraging early results on the use of deep learning to compute ensembles of macromolecular structures. Critically for the usability of computed structures, for both single proteins and protein complexes, deep learning derived estimates of both local and global accuracy are of high quality, however the estimates in interface regions are slightly less reliable. CASP15 also included computation of RNA structures for the first time. Here, the classical approaches produced better agreement with experiment than the new deep learning ones, and accuracy is limited. Also, for the first time, CASP included the computation of protein-ligand complexes, an area of special interest for drug design. Here too, classical methods were still superior to deep learning ones. Many new approaches were discussed at the CASP conference, and it is clear methods will continue to advance.

CASP16 Protein Monomer Structure Prediction Assessment. Yuan et al. (2025). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.70031

The assessment of monomer targets in the Critical Assessment of Structure Prediction Round 16 (CASP16) underscores that the problem of single-domain protein fold prediction is nearly solved-no target folds were incorrectly predicted across all Evaluation Units. However, challenges remain in accurately modeling truncated sequences, irregular secondary structures, and interaction-induced conformational changes. The release of AlphaFold3 (AF3) during CASP16, and its effective integration by many groups, demonstrated its superiority over AlphaFold2 (AF2), particularly in confidence estimation and model selection. Additional improvements in multiple sequence alignments (MSAs) and fragment-based prediction, that is, selecting the optimal fragment of the full sequence for modeling, also contributed to enhanced prediction accuracy. The top three groups-all from the Yang lab-consistently outperformed others across CASP16 monomer targets, reflecting their robust modeling pipelines and successful adoption of AF3. CASP16 also introduced three new challenges: Phase 0, in which stoichiometry was withheld; Phase 2, which supplied ~8000 MassiveFold models per target to test model selection strategies; and Model 6, which limited predictors to using MSAs provided by the organizers. While we evaluated group performance in these additional challenges, the insights gained were limited due to low participation and caveats in the design of experiments. We suggest improvements for the organization of these challenges and encourage broader engagement from the prediction community. The progress in monomer modeling from CASP15 to CASP16 was subtle, but more groups in CASP16 were able to outperform ColabFold, reflecting the community's improved ability in optimizing AF2 and the growing adoption of AF3. We anticipate that the recent release of the AF3 source code will stimulate future progress through user-driven optimization and innovations in model architecture. Finally, model ranking remains a persistent weakness across most groups, highlighting a critical area for future development.

- Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Silver, et al. (2021) — AlphaFold2 at CASP14.

Applying and improving AlphaFold at CASP14. Jumper et al. (2021). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.26257

We describe the operation and improvement of AlphaFold, the system that was entered by the team AlphaFold2 to the ”human” category in the 14th Critical Assessment of Protein Structure Prediction (CASP14). The AlphaFold system entered in CASP14 is entirely different to the one entered in CASP13. It used a novel end-to-end deep neural network trained to produce protein structures from amino acid sequence, multiple sequence alignments, and homologous proteins. In the assessors' ranking by summed z scores (>2.0), AlphaFold scored 244.0 compared to 90.8 by the next best group. The predictions made by AlphaFold had a median domain GDT_TS of 92.4; this is the first time that this level of average accuracy has been achieved during CASP, especially on the more difficult Free Modeling targets, and represents a significant improvement in the state of the art in protein structure prediction. We reported how AlphaFold

was run as a human team during CASP14 and improved such that it now achieves an equivalent level of performance without intervention, opening the door to highly accurate large-scale structure prediction.

- Zemla (2003), Zhang and Skolnick (2004), Mariani et al. (2013), Basu and Wallner (2016) — GDT, TM-score, lDDT and DockQ.

LGA: a method for finding 3D similarities in protein structures. Zemla (2003). *Nucleic Acids Research*. doi:10.1093/nar/gkg571

> We present the LGA (Local-Global Alignment) method, designed to facilitate the comparison of protein structures or fragments of protein structures in sequence dependent and sequence independent modes. The LGA structure alignment program is available as an online service at http://PredictionCenter.llnl.gov/local/lga. Data generated by LGA can be successfully used in a scoring function to rank the level of similarity between two structures and to allow structure classification when many proteins are being analyzed. LGA also allows the clustering of similar fragments of protein structures.

Scoring function for automated assessment of protein structure template quality. Zhang & Skolnick (2004). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.20264

> We have developed a new scoring function, the template modeling score (TM-score), to assess the quality of protein structure templates and predicted full-length models by extending the approaches used in Global Distance Test (GDT)1 and MaxSub.2 First, a protein size-dependent scale is exploited to eliminate the inherent protein size dependence of the previous scores and appropriately account for random protein structure pairs. Second, rather than setting specific distance cutoffs and calculating only the fractions with errors below the cutoff, all residue pairs in alignment/modeling are evaluated in the proposed score. For comparison of various scoring functions, we have constructed a large-scale benchmark set of structure templates for 1489 small to medium size proteins using the threading program PROSPECTOR_3 and built the full-length models using MODELLER and TASSER. The TM-score of the initial threading alignments, compared to the GDT and MaxSub scoring functions, shows a much stronger correlation to the quality of the final full-length models. The TM-score is further exploited as an assessment of all 'new fold' targets in the recent CASP5 experiment and shows a close coincidence with the results of human-expert visual assessment. These data suggest that the TM-score is a useful complement to the fully automated assessment of protein structure predictions. The executable program of TM-score is freely downloadable at http://bioinformatics.buffalo.edu/TM-score.

lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Mariani et al. (2013). *Bioinformatics*. doi:10.1093/bioinformatics/btt473

> Motivation The assessment of protein structure prediction techniques requires objective criteria to measure the similarity between a computational model and the experimentally determined reference structure. Conventional similarity measures based on a global superposition of carbon α atoms are strongly influenced by domain motions and do not assess the accuracy of local atomic details in the model. Results The Local Distance Difference Test (lDDT) is a superposition-free score that evaluates local distance differences of all atoms in a model, including validation of stereochemical plausibility. The reference can be a single structure, or an ensemble of equivalent structures. We demonstrate that lDDT is well suited to assess local model quality, even in the presence of domain movements, while maintaining good correlation with global measures. These properties make lDDT a robust tool for the automated assessment of structure prediction servers without manual intervention. Availability and implementation Source code, binaries for Linux and MacOSX, and an interactive web server are available at http://swissmodel.expasy.org/lddt. Contact torsten.schwede@unibas.ch. Supplementary information Supplementary data are available at Bioinformatics online.

DockQ: A Quality Measure for Protein-Protein Docking Models. Basu & Wallner (2016). *PLOS ONE*. doi:10.1371/journal.pone.0161879

> The state-of-the-art to assess the structural quality of docking models is currently based on three related yet independent quality measures: Fnat, LRMS, and iRMS as proposed and standardized by CAPRI. These quality measures quantify different aspects of the quality of a particular docking model and need to be viewed together to reveal the true quality, e.g. a model with relatively poor LRMS (>10Å) might still qualify as 'acceptable' with a descent Fnat (>0.50) and iRMS (<3.0Å). This is also the reason why the so called CAPRI criteria for assessing the quality of docking models is defined by applying various ad-hoc cutoffs on these measures to classify a docking model into the four classes: Incorrect, Acceptable, Medium,

or High quality. This classification has been useful in CAPRI, but since models are grouped in only four bins it is also rather limiting, making it difficult to rank models, correlate with scoring functions or use it as target function in machine learning algorithms. Here, we present DockQ, a continuous protein-protein docking model quality measure derived by combining Fnat, LRMS, and iRMS to a single score in the range [0, 1] that can be used to assess the quality of protein docking models. By using DockQ on CAPRI models it is possible to almost completely reproduce the original CAPRI classification into Incorrect, Acceptable, Medium and High quality. An average PPV of 94% at 90% Recall demonstrating that there is no need to apply predefined ad-hoc cutoffs to classify docking models. Since DockQ recapitulates the CAPRI classification almost perfectly, it can be viewed as a higher resolution version of the CAPRI classification, making it possible to estimate model quality in a more quantitative way using Z-scores or sum of top ranked models, which has been so valuable for the CASP community. The possibility to directly correlate a quality measure to a scoring function has been crucial for the development of scoring functions for protein structure prediction, and DockQ should be useful in a similar development in the protein docking field. DockQ is available at http://github.com/bjornwallner/DockQ/.

- Janin et al. (2003), Haas et al. (2017) — CAPRI and CAMEO.

  CAPRI: A Critical Assessment of PRedicted Interactions. Janin et al. (2003). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.10381

  > CAPRI is a communitywide experiment to assess the capacity of protein-docking methods to predict protein-protein interactions. Nineteen groups participated in rounds 1 and 2 of CAPRI and submitted blind structure predictions for seven protein-protein complexes based on the known structure of the component proteins. The predictions were compared to the unpublished X-ray structures of the complexes. We describe here the motivations for launching CAPRI, the rules that we applied to select targets and run the experiment, and some conclusions that can already be drawn. The results stress the need for new scoring functions and for methods handling the conformation changes that were observed in some of the target systems. CAPRI has already been a powerful drive for the community of computational biologists who development docking algorithms. We hope that this issue of Proteins will also be of interest to the community of structural biologists, which we call upon to provide new targets for future rounds of CAPRI, and to all molecular biologists who view protein-protein recognition as an essential process.

  Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Haas et al. (2017). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.25431

  > Every second year, the community experiment "Critical Assessment of Techniques for Structure Prediction" (CASP) is conducting an independent blind assessment of structure prediction methods, providing a framework for comparing the performance of different approaches and discussing the latest developments in the field. Yet, developers of automated computational modeling methods clearly benefit from more frequent evaluations based on larger sets of data. The "Continuous Automated Model EvaluatiOn (CAMEO)" platform complements the CASP experiment by conducting fully automated blind prediction assessments based on the weekly pre-release of sequences of those structures, which are going to be published in the next release of the PDB Protein Data Bank. CAMEO publishes weekly benchmarking results based on models collected during a 4-day prediction window, on average assessing ca. 100 targets during a time frame of 5 weeks. CAMEO benchmarking data is generated consistently for all participating methods at the same point in time, enabling developers to benchmark and cross-validate their method's performance, and directly refer to the benchmarking results in publications. In order to facilitate server development and promote shorter release cycles, CAMEO sends weekly email with submission statistics and low performance warnings. Many participants of CASP have successfully employed CAMEO when preparing their methods for upcoming community experiments. CAMEO offers a variety of scores to allow benchmarking diverse aspects of structure prediction methods. By introducing new scoring schemes, CAMEO facilitates new development in areas of active research, for example, modeling quaternary structure, complexes, or ligand binding sites.

B.9.2 Structure and its determination

- Pauling et al. (1951), Ramachandran et al. (1963) — the α-helix and the Ramachandran plot.

The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Pauling et al. (1951). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.37.4.205

Stereochemistry of polypeptide chain configurations. Ramachandran et al. (1963). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(63)80023-6

- Kendrew et al. (1958), PERUTZ et al. (1960) — the first protein structures.

A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Kendrew et al. (1958). *Nature*. doi:10.1038/181662a0

Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. PERUTZ et al. (1960). *Nature*. doi:10.1038/185416a0

- Chothia et al. (1977), 2, Richardson (1981) — packing, structural classes and the anatomy of proteins.

Structure of proteins: packing of alpha-helices and pleated sheets.. Chothia et al. (1977). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.74.10.4130

> Simple models are presented that describe the rules for almost all the packing that occurs between and among alpha-helices and pleated sheets. These packing rules, together with the primary and secondary structures, are the major determinants of the three-dimensional structure of proteins.

The Anatomy and Taxonomy of Protein Structure. Richardson (1981). *Advances in Protein Chemistry Volume 34*. doi:10.1016/s0065-3233(08)60520-3

- Williamson et al. (1985) — the first NMR structure of a protein.

Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. Williamson et al. (1985). *Journal of Molecular Biology*. doi:10.1016/0022-2836(85)90347-x

> A determination of the solution conformation of the proteinase inhibitor IIA from bull seminal plasma (BUSI IIA) is described. Two-dimensional nuclear Overhauser enhancement spectroscopy (NOESY) was used to obtain a list of 202 distance constraints between individually assigned hydrogen atoms of the polypeptide chain, to identify the positions of the three disulfide bridges, and to locate the single cis peptide bond. Supplementary geometric constraints were derived from the vicinal spin-spin couplings and the locations of certain hydrogen bonds, as determined by nuclear magnetic resonance (n.m.r.). Using a new distance geometry program (DISGEO) which is capable of computing all-atom structures for proteins the size of BUSI IIA, five conformers were computed from the NOE distance constraints alone, and another five were computed with the supplementary constraints included. Comparison of the different structures computed from the n.m.r. data among themselves and with the crystal structures of two homologous proteins shows that the global features of the conformation of BUSI IIA (i.e. the overall dimensions of the molecule and the threading of the polypeptide chain) were well-defined by the available n.m.r. data. In the Appendix, we describe a preliminary energy refinement of the structure, which showed that the constraints derived from the n.m.r. data are compatible with a low energy spatial structure.

- Brünger (1992) — the free R value.

Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Brünger (1992). *Nature*. doi:10.1038/355472a0

> The determination of macromolecular structure by crystallography involves fitting atomic models to the observed diffraction data. The traditional measure of the quality of this fit, and presumably the accuracy of the model, is the R value. Despite stereochemical restraints, it is possible to overfit or 'misfit' the diffraction data: an incorrect model can be refined to fairly good R values as several recent examples have shown. Here I propose a reliable and unbiased indicator of the accuracy of such models. By analogy with the cross-validation method of testing statistical models I define a statistical quantity (R(free) (T) that measures the agreement between observed and computed structure factor amplitudes for a 'test' set of reflections that is omitted in the modelling and refinement process. As examples show, there is a

high correlation between R(free) (T) and the accuracy of the atomic model phases. This is useful because experimental phase information is usually inaccurate, incomplete or unavailable. I expect that R(free) (T) will provide a measure of the information content of recently proposed models of thermal motion and disorder, time-averaging and bulk solvent.

- Henderson et al. (1990), Kühlbrandt (2014), Scheres (2012) — cryo-EM and the resolution revolution.

Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. Henderson et al. (1990). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(05)80271-2

> The light-driven proton pump bacteriorhodopsin occurs naturally as two-dimensional crystals. A three-dimensional density map of the structure, at near-atomic resolution, has been obtained by studying the crystals using electron cryo-microscopy to obtain electron diffraction patterns and high-resolution micrographs. New methods were developed for analysing micrographs from tilted specimens, incorporating methods previously developed for untilted specimens that enable large areas to be analysed and corrected for distortions. Data from 72 images, from both tilted and untilted specimens, were analysed to produce the phases of 2700 independent Fourier components of the structure. The amplitudes of these components were accurately measured from 150 diffraction patterns. Together, these data represent about half of the full three-dimensional transform to 3.5 A. The map of the structure has a resolution of 3.5 A in a direction parallel to the membrane plane but lower than this in the perpendicular direction. It shows many features in the density that are resolved from the main density of the seven alpha-helices. We interpret these features as the bulky aromatic side-chains of phenylalanine, tyrosine and tryptophan residues. There is also a very dense feature, which is the beta-ionone ring of the retinal chromophore. Using these bulky side-chains as guide points and taking account of bulges in the helices that indicate smaller side-chains such as leucine, a complete atomic model for bacteriorhodopsin between amino acid residues 8 and 225 has been built. There are 21 amino acid residues, contributed by all seven helices, surrounding the retinal and 26 residues, contributed by five helices, forming the proton pathway or channel. Ten of the amino acid residues in the middle of the proton channel are also part of the retinal binding site. The model also provides a useful basis for consideration of the mechanism of proton pumping and allows a consistent interpretation of a great deal of other experimental data. In particular, the structure suggests that pK changes in the Schiff base must act as the means by which light energy is converted into proton pumping pressure in the channel. Asp96 is on the pathway from the cytoplasm to the Schiff base and Asp85 is on the pathway from the Schiff base to the extracellular surface.

The Resolution Revolution. Kühlbrandt (2014). *Science*. doi:10.1126/science.1251652

> Advances in detector technology and image processing are yielding high-resolution electron cryo-microscopy structures of biomolecules. [Also see Report by Amunts et al.]

A Bayesian View on Cryo-EM Structure Determination. Scheres (2012). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2011.11.010

> Three-dimensional (3D) structure determination by single-particle analysis of cryo-electron microscopy (cryo-EM) images requires many parameters to be determined from extremely noisy data. This makes the method prone to overfitting, that is, when structures describe noise rather than signal, in particular near their resolution limit where noise levels are highest. Cryo-EM structures are typically filtered using ad hoc procedures to prevent overfitting, but the tuning of arbitrary parameters may lead to subjectivity in the results. I describe a Bayesian interpretation of cryo-EM structure determination, where smoothness in the reconstructed density is imposed through a Gaussian prior in the Fourier domain. The statistical framework dictates how data and prior knowledge should be combined, so that the optimal 3D linear filter is obtained without the need for arbitrariness and objective resolution estimates may be obtained. Application to experimental data indicates that the statistical approach yields more reliable structures than existing methods and is capable of detecting smaller classes in data sets that contain multiple different structures.

B.9.3 Folding

- Levinthal (1966) — Levinthal's paradox.

Molecular Model-Building by Computer. Levinthal (1966). *Scientific American*. doi:10.1038/scientificamerican0666-42

- Jackson and Fersht (1991) — two-state folding.

  Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Jackson & Fersht (1991). *Biochemistry*. doi:10.1021/bi00107a010

  > The reversible folding and unfolding of barley chymotrypsin inhibitor 2 (CI2) appears to be a rare example in which both equilibria and kinetics are described by a two-state model. Equilibrium denaturation by guanidinium chloride and heat is completely reversible, and the data can be fitted to a simple two-state model involving only native and denatured forms. The free energy of folding in the absence of denaturant, delta GH2O, at pH 6.3, is calculated to be 7.03 +/- 0.16 and 7.18 +/- 0.43 kcal mol-1 for guanidinium chloride and thermal denaturation, respectively. Scanning microcalorimetry shows that the ratio of the van't Hoff enthalpy of denaturation to the calorimetric enthalpy of denaturation does not deviate from unity, the value observed for a two-state transition, over the pH range 2.2-3.5. The heat capacity change for denaturation is found to be 0.789 kcal mol-1 K-1. The rate of unfolding of CI2 is first order and increases exponentially with increasing guanidinium chloride concentration. Refolding, however, is complex and involves at least three well-resolved phases. The three phases result from heterogeneity of the unfolded form due to proline isomerization. The fast phase, 77% of the amplitude, corresponds to the refolding of the fraction of the protein that has all its prolines in a native trans conformation. The rate of this major phase decreases exponentially with increasing guanidinium chloride concentration. The unfolding and refolding kinetics can also be fitted to a two-state model.(ABSTRACT TRUNCATED AT 250 WORDS)

- Bryngelson et al. (1995), Dill and Chan (1997) — energy landscapes and funnels.

  Funnels, pathways, and the energy landscape of protein folding: A synthesis. Bryngelson et al. (1995). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.340210302

  > The understanding, and even the description of protein folding is impeded by the complexity of the process. Much of this complexity can be described and understood by taking a statistical approach to the energetics of protein conformation, that is, to the energy landscape. The statistical energy landscape approach explains when and why unique behaviors, such as specific folding pathways, occur in some proteins and more generally explains the distinction between folding processes common to all sequences and those peculiar to individual sequences. This approach also gives new, quantitative insights into the interpretation of experiments and simulations of protein folding thermodynamics and kinetics. Specifically, the picture provides simple explanations for folding as a two-state first-order phase transition, for the origin of metastable collapsed unfolded states and for the curved Arrhenius plots observed in both laboratory experiments and discrete lattice simulations. The relation of these quantitative ideas to folding pathways, to uniexponential vs. multiexponential behavior in protein folding experiments and to the effect of mutations on folding is also discussed. The success of energy landscape ideas in protein structure prediction is also described. The use of the energy landscape approach for analyzing data is illustrated with a quantitative analysis of some recent simulations, and a qualitative analysis of experiments on the folding of three proteins. The work unifies several previously proposed ideas concerning the mechanism protein folding and delimits the regions of validity of these ideas under different thermodynamic conditions.

  From Levinthal to pathways to funnels. Dill & Chan (1997). *Nature Structural & Molecular Biology*. doi:10.1038/nsb0197-10

  > While the classical view of protein folding kinetics relies on phenomenological models, and regards folding intermediates in a structural way, the new view emphasizes the ensemble nature of protein conformations. Although folding has sometimes been regarded as a linear sequence of events, the new view sees folding as parallel microscopic multi-pathway diffusion-like processes. While the classical view invoked pathways to solve the problem of searching for the needle in the haystack, the pathway idea was then seen as conflicting with Anfinsen's experiments showing that folding is pathway-independent (Levinthal's paradox). In contrast, the new view sees no inherent paradox because it eliminates the pathway idea: folding can funnel to a single stable state by multiple routes in conformational space. The general energy landscape picture provides a conceptual framework for understanding both two-state and multi-state folding kinetics. Better tests of these ideas will come when new experiments become available for measuring not just averages of structural observables, but also correlations among their fluctuations. At that point we hope to learn much more about the real shapes of protein folding landscapes.

- Dill and MacCallum (2012) — the protein-folding problem, 50 years on.

  The Protein-Folding Problem, 50 Years On. Dill & MacCallum (2012). *Science*. doi:10.1126/science.1219021

  > The protein-folding problem was first posed about one half-century ago. The term refers to three broad questions: (i) What is the physical code by which an amino acid sequence dictates a protein's native structure? (ii) How can proteins fold so fast? (iii) Can we devise a computer algorithm to predict protein structures from their sequences? We review progress on these problems. In a few cases, computer simulations of the physical forces in chemically detailed models have now achieved the accurate folding of small proteins. We have learned that proteins fold rapidly because random thermal motions cause conformational changes leading energetically downhill toward the native structure, a principle that is captured in funnel-shaped energy landscapes. And thanks in part to the large Protein Data Bank of known structures, predicting protein structures is now far more successful than was thought possible in the early days. What began as three questions of basic science one half-century ago has now grown into the full-fledged research field of protein physical science.

- Shaw et al. (2010), Lindorff-Larsen et al. (2011) — folding by molecular dynamics.

  Atomic-Level Characterization of the Structural Dynamics of Proteins. Shaw et al. (2010). *Science*. doi:10.1126/science.1187409

  > Molecular dynamics (MD) simulations are widely used to study protein motions at an atomic level of detail, but they have been limited to time scales shorter than those of many biologically critical conformational changes. We examined two fundamental processes in protein dynamics--protein folding and conformational change within the folded state--by means of extremely long all-atom MD simulations conducted on a special-purpose machine. Equilibrium simulations of a WW protein domain captured multiple folding and unfolding events that consistently follow a well-defined folding pathway; separate simulations of the protein's constituent substructures shed light on possible determinants of this pathway. A 1-millisecond simulation of the folded protein BPTI reveals a small number of structurally distinct conformational states whose reversible interconversion is slower than local relaxations within those states by a factor of more than 1000.

  How Fast-Folding Proteins Fold. Lindorff-Larsen et al. (2011). *Science*. doi:10.1126/science.1208351

  > An outstanding challenge in the field of molecular biology has been to understand the process by which proteins fold into their characteristic three-dimensional structures. Here, we report the results of atomic-level molecular dynamics simulations, over periods ranging between 100 μs and 1 ms, that reveal a set of common principles underlying the folding of 12 structurally diverse proteins. In simulations conducted with a single physics-based energy function, the proteins, representing all three major structural classes, spontaneously and repeatedly fold to their experimentally determined native structures. Early in the folding process, the protein backbone adopts a nativelike topology while certain secondary structure elements and a small number of nonlocal contacts form. In most cases, folding follows a single dominant route in which elements of the native structure appear in an order highly correlated with their propensity to form in the unfolded state.

B.9.4 Comparison and classification

- Chothia and Lesk (1986) — structure is more conserved than sequence.

  The relation between the divergence of sequence and structure in proteins.. Chothia & Lesk (1986). *The EMBO Journal*. doi:10.1002/j.1460-2075.1986.tb04288.x

  > Homologous proteins have regions which retain the same general fold and regions where the folds differ. For pairs of distantly related proteins (residue identity approximately 20%), the regions with the same fold may comprise less than half of each molecule. The regions with the same general fold differ in structure by amounts that increase as the amino acid sequences diverge. The root mean square deviation in the positions of the main chain atoms, delta, is related to the fraction of mutated residues, H, by the expression: delta(A) = 0.40 e1.87H.

- Holm and Sander (1993), Holm and Sander (1996), Holm (2019) — DALI and the map of the protein universe.

  Protein Structure Comparison by Alignment of Distance Matrices. Holm & Sander (1993). *Journal of Molecular Biology*. doi:10.1006/jmbi.1993.1489

  > With a rapidly growing pool of known tertiary structures, the importance of protein structure comparison parallels that of sequence alignment. We have developed a novel algorithm (DALI) for optimal pairwise alignment of protein structures. The three-dimensional co-ordinates of each protein are used to calculate residue-residue (C alpha-C alpha) distance matrices. The distance matrices are first decomposed into elementary contact patterns, e.g. hexapeptide-hexapeptide submatrices. Then, similar contact patterns in the two matrices are paired and combined into larger consistent sets of pairs. A Monte Carlo procedure is used to optimize a similarity score defined in terms of equivalent intramolecular distances. Several alignments are optimized in parallel, leading to simultaneous detection of the best, second-best and so on solutions. The method allows sequence gaps of any length, reversal of chain direction and free topological connectivity of aligned segments. Sequential connectivity can be imposed as an option. The method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, even in the presence of geometrical distortions. An all-against-all alignment of over 200 representative protein structures results in an objective classification of known three-dimensional folds in agreement with visual classifications. Unexpected topological similarities of biological interest have been detected, e.g. between the bacterial toxin colicin A and globins, and between the eukaryotic POU-specific DNA-binding domain and the bacterial lambda repressor.

  Mapping the Protein Universe. Holm & Sander (1996). *Science*. doi:10.1126/science.273.5275.595

  > The comparison of the three-dimensional shapes of protein molecules poses a complex algorithmic problem. Its solution provides biologists with computational tools to organize the rapidly growing set of thousands of known protein shapes, to identify new types of protein architecture, and to discover unexpected evolutionary relations, reaching back billions of years, between protein molecules. Protein shape comparison also improves tools for identifying gene functions in genome databases by defining the essential sequence-structure features of a protein family. Finally, an exhaustive all-on-all shape comparison provides a map of physical attractor regions in the abstract shape space of proteins, with implications for the processes of protein folding and evolution.

  DALI and the persistence of protein shape. Holm (2019). *Protein Science*. doi:10.1002/pro.3749

  > DALI is a popular resource for comparing protein structures. The software is based on distance-matrix alignment. The associated web server provides tools to navigate, integrate and organize some data pushed out by genomics and structural genomics. The server has been running continuously for the past 25 years. Structural biologists routinely use DALI to compare a new structure against previously known protein structures. If significant similarities are discovered, it may indicate a distant homology, that is, that the structures are of shared origin. This may be significant in determining the molecular mechanisms, as these may remain very similar from a distant predecessor to the present day, for example, from the last common ancestor of humans and bacteria. Meta-analysis of independent reference-based evaluations of alignment accuracy and fold discrimination shows DALI at top rank in six out of 12 studies. The web server and standalone software are available from http://ekhidna2.biocenter.helsinki.fi/dali.

- Zhang and Skolnick (2005), Kolodny et al. (2005) — TM-align and the evaluation of structural aligners.

  TM-align: a protein structure alignment algorithm based on the TM-score. Zhang & Skolnick (2005). *Nucleic Acids Research*. doi:10.1093/nar/gki524

  > We have developed TM-align, a new algorithm to identify the best structural alignment between protein pairs that combines the TM-score rotation matrix and Dynamic Programming (DP). The algorithm is approximately 4 times faster than CE and 20 times faster than DALI and SAL. On average, the resulting structure alignments have higher accuracy and coverage than those provided by these most often-used methods. TM-align is applied to an all-against-all structure comparison of 10 515 representative protein chains from the Protein Data Bank (PDB) with a sequence identity cutoff <95%: 1996 distinct folds are found when a TM-score threshold of 0.5 is used. We also use TM-align to match the models predicted by TASSER for solved non-homologous proteins in PDB. For both folded and misfolded models, TM-align can almost always find close structural analogs, with an average root mean square deviation, RMSD, of 3 A and 87% alignment coverage. Nevertheless, there exists a significant correlation between the correctness

of the predicted structure and the structural similarity of the model to the other proteins in the PDB. This correlation could be used to assist in model selection in blind protein structure predictions. The TM-align program is freely downloadable at http://bioinformatics.buffalo.edu/TM-align.

Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures. Kolodny et al. (2005). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2004.12.032

> We report the largest and most comprehensive comparison of protein structural alignment methods. Specifically, we evaluate six publicly available structure alignment programs: SSAP, STRUCTAL, DALI, LSQMAN, CE and SSM by aligning all 8,581,970 protein structure pairs in a test set of 2930 protein domains specially selected from CATH v.2.4 to ensure sequence diversity. We consider an alignment good if it matches many residues, and the two substructures are geometrically similar. Even with this definition, evaluating structural alignment methods is not straightforward. At first, we compared the rates of true and false positives using receiver operating characteristic (ROC) curves with the CATH classification taken as a gold standard. This proved unsatisfactory in that the quality of the alignments is not taken into account: sometimes a method that finds less good alignments scores better than a method that finds better alignments. We correct this intrinsic limitation by using four different geometric match measures (SI, MI, SAS, and GSAS) to evaluate the quality of each structural alignment. With this improved analysis we show that there is a wide variation in the performance of different methods; the main reason for this is that it can be difficult to find a good structural alignment between two proteins even when such an alignment exists. We find that STRUCTAL and SSM perform best, followed by LSQMAN and CE. Our focus on the intrinsic quality of each alignment allows us to propose a new method, called "Best-of-All" that combines the best results of all methods. Many commonly used methods miss 10-50% of the good Best-of-All alignments. By putting existing structural alignments into proper perspective, our study allows better comparison of protein structures. By highlighting limitations of existing methods, it will spur the further development of better structural alignment methods. This will have significant biological implications now that structural comparison has come to play a central role in the analysis of experimental work on protein structure, protein function and protein evolution.

- van Kempen et al. (2023), Barrio-Hernandez et al. (2023) — Foldseek and the clustering of the AlphaFold Database.

  Fast and accurate protein structure search with Foldseek. van Kempen et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-023-01773-0

  > As structure prediction methods are generating millions of publicly available protein structures, searching these databases is becoming a bottleneck. Foldseek aligns the structure of a query protein against a database by describing tertiary amino acid interactions within proteins as sequences over a structural alphabet. Foldseek decreases computation times by four to five orders of magnitude with 86%, 88% and 133% of the sensitivities of Dali, TM-align and CE, respectively.

  Clustering predicted structures at the scale of the known protein universe. Barrio-Hernandez et al. (2023). *Nature*. doi:10.1038/s41586-023-06510-w

  > Proteins are key to all cellular processes and their structure is important in understanding their function and evolution. Sequence-based predictions of protein structures have increased in accuracy 1 , and over 214 million predicted structures are available in the AlphaFold database 2 . However, studying protein structures at this scale requires highly efficient methods. Here, we developed a structural-alignment-based clustering algorithm-Foldseek cluster-that can cluster hundreds of millions of structures. Using this method, we have clustered all of the structures in the AlphaFold database, identifying 2.30 million non-singleton structural clusters, of which 31% lack annotations representing probable previously undescribed structures. Clusters without annotation tend to have few representatives covering only 4% of all proteins in the AlphaFold database. Evolutionary analysis suggests that most clusters are ancient in origin but 4% seem to be species specific, representing lower-quality predictions or examples of de novo gene birth. We also show how structural comparisons can be used to predict domain families and their relationships, identifying examples of remote structural similarity. On the basis of these analyses, we identify several examples of human immune-related proteins with putative remote homology in prokaryotic species, illustrating the value of this resource for studying protein function and evolution across the tree of life.

- Chothia (1992) — one thousand families.

  One thousand families for the molecular biologist. Chothia (1992). *Nature*. doi:10.1038/357543a0

- Murzin et al. (1995), Sillitoe et al. (2020), Cheng et al. (2014), Lau et al. (2024) — SCOP, CATH, ECOD and TED.

SCOP: A structural classification of proteins database for the investigation of sequences and structures. Murzin et al. (1995). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(05)80134-2

CATH: increased structural coverage of functional space. Sillitoe et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa1079

> CATH (https://www.cathdb.info) identifies domains in protein structures from wwPDB and classifies these into evolutionary superfamilies, thereby providing structural and functional annotations. There are two levels: CATH-B, a daily snapshot of the latest domain structures and superfamily assignments, and CATH+, with additional derived data, such as predicted sequence domains, and functionally coherent sequence subsets (Functional Families or FunFams). The latest CATH+ release, version 4.3, significantly increases coverage of structural and sequence data, with an addition of 65,351 fully-classified domains structures (+15%), providing 500 238 structural domains, and 151 million predicted sequence domains (+59%) assigned to 5481 superfamilies. The FunFam generation pipeline has been re-engineered to cope with the increased influx of data. Three times more sequences are captured in FunFams, with a concomitant increase in functional purity, information content and structural coverage. FunFam expansion increases the structural annotations provided for experimental GO terms (+59%). We also present CATH-FunVar web-pages displaying variations in protein sequences and their proximity to known or predicted functional sites. We present two case studies (1) putative cancer drivers and (2) SARS-CoV-2 proteins. Finally, we have improved links to and from CATH including SCOP, InterPro, Aquaria and 2DProt.

ECOD: An Evolutionary Classification of Protein Domains. Cheng et al. (2014). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1003926

> Understanding the evolution of a protein, including both close and distant relationships, often reveals insight into its structure and function. Fast and easy access to such up-to-date information facilitates research. We have developed a hierarchical evolutionary classification of all proteins with experimentally determined spatial structures, and presented it as an interactive and updatable online database. ECOD (Evolutionary Classification of protein Domains) is distinct from other structural classifications in that it groups domains primarily by evolutionary relationships (homology), rather than topology (or "fold"). This distinction highlights cases of homology between domains of differing topology to aid in understanding of protein structure evolution. ECOD uniquely emphasizes distantly related homologs that are difficult to detect, and thus catalogs the largest number of evolutionary links among structural domain classifications. Placing distant homologs together underscores the ancestral similarities of these proteins and draws attention to the most important regions of sequence and structure, as well as conserved functional sites. ECOD also recognizes closer sequence-based relationships between protein domains. Currently, approximately 100,000 protein structures are classified in ECOD into 9,000 sequence families clustered into close to 2,000 evolutionary groups. The classification is assisted by an automated pipeline that quickly and consistently classifies weekly releases of PDB structures and allows for continual updates. This synchronization with PDB uniquely distinguishes ECOD among all protein classifications. Finally, we present several case studies of homologous proteins not recorded in other classifications, illustrating the potential of how ECOD can be used to further biological and evolutionary studies.

Exploring structural diversity across the protein universe with The Encyclopedia of Domains. Lau et al. (2024). *Science*. doi:10.1126/science.adq4946

> The AlphaFold Protein Structure Database (AFDB) contains more than 214 million predicted protein structures composed of domains, which are independently folding units found in multiple structural and functional contexts. Identifying domains can enable many functional and evolutionary analyses but has remained challenging because of the sheer scale of the data. Using deep learning methods, we have detected and classified every domain in the AFDB, producing The Encyclopedia of Domains. We detected nearly 365 million domains, over 100 million more than can be found by sequence methods, covering more than 1 million taxa. Reassuringly, 77% of the nonredundant domains are similar to known superfamilies, greatly expanding representation of their domain space. We uncovered more than 10,000 new structural interactions between superfamilies and thousands of new folds across the fold space continuum.

- Varadi et al. (2023), Bertoni et al. (2025) — the AlphaFold Database.

  AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Varadi et al. (2023). *Nucleic Acids Research*. doi:10.1093/nar/gkad1011

  > The AlphaFold Database Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk) has significantly impacted structural biology by amassing over 214 million predicted protein structures, expanding from the initial 300k structures released in 2021. Enabled by the groundbreaking AlphaFold2 artificial intelligence (AI) system, the predictions archived in AlphaFold DB have been integrated into primary data resources such as PDB, UniProt, Ensembl, InterPro and MobiDB. Our manuscript details subsequent enhancements in data archiving, covering successive releases encompassing model organisms, global health proteomes, Swiss-Prot integration, and a host of curated protein datasets. We detail the data access mechanisms of AlphaFold DB, from direct file access via FTP to advanced queries using Google Cloud Public Datasets and the programmatic access endpoints of the database. We also discuss the improvements and services added since its initial release, including enhancements to the Predicted Aligned Error viewer, customisation options for the 3D viewer, and improvements in the search engine of AlphaFold DB.

  AlphaFold Protein Structure Database 2025: a redesigned interface and updated structural coverage. Bertoni et al. (2025). *Nucleic Acids Research*. doi:10.1093/nar/gkaf1226

  > The AlphaFold Protein Structure Database (AFDB; https://alphafold.ebi.ac.uk), developed by EMBL-EBI and Google DeepMind, provides open access to hundreds of millions of high-accuracy protein structure predictions, transforming research in structural biology and the wider life sciences. Since its launch, AFDB has become a widely used bioinformatics resource, integrated into major databases, visualization platforms, and analysis pipelines. Here, we report the update of the database to align with the UniProt 2025_03 release, along with a comprehensive redesign of the entry page to enhance usability, accessibility, and structural interpretation. The new design integrates annotations directly with an interactive 3D viewer and introduces dedicated domains and summary tabs. Structural coverage has also been updated to include isoforms plus underlying multiple sequence alignments. Data are available through the website, FTP, Google Cloud, and updated APIs. Together, these advances reinforce AFDB as a sustainable resource for exploring protein sequence-structure relationships.

B.9.5 Structure prediction before deep learning

- Chou and Fasman (1974a), Chou and Fasman (1974b), Garnier et al. (1978) — early secondary structure prediction.

  Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Chou & Fasman (1974). *Biochemistry*. doi:10.1021/bi00699a001

  Prediction of protein conformation. Chou & Fasman (1974). *Biochemistry*. doi:10.1021/bi00699a002

  Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Garnier et al. (1978). *Journal of Molecular Biology*. doi:10.1016/0022-2836(78)90297-8

- ⋆ Rost and Sander (1993), Jones (1999) — PHD and PSIPRED.

  Prediction of Protein Secondary Structure at Better than 70% Accuracy. Rost & Sander (1993). *Journal of Molecular Biology*. doi:10.1006/jmbi.1993.1413

  > We have trained a two-layered feed-forward neural network on a non-redundant data base of 130 protein chains to predict the secondary structure of water-soluble proteins. A new key aspect is the use of evolutionary information in the form of multiple sequence alignments that are used as input in place of single sequences. The inclusion of protein family information in this form increases the prediction accuracy by six to eight percentage points. A combination of three levels of networks results in an overall three-state accuracy of 70.8% for globular proteins (sustained performance). If four membrane protein chains are included in the evaluation, the overall accuracy drops to 70.2%. The prediction is well balanced between alpha-helix, beta-strand and loop: 65% of the observed strand residues are predicted correctly.

> The accuracy in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy. The performance accuracy is verified by a sevenfold cross-validation test, and an additional test on 26 recently solved proteins. Of particular practical importance is the definition of a position-specific reliability index. For half of the residues predicted with a high level of reliability the overall accuracy increases to better than 82%. A further strength of the method is the more realistic prediction of segment length. The protein family prediction method is available for testing by academic researchers via an electronic mail server.

Protein secondary structure prediction based on position-specific scoring matrices. Jones (1999). *Journal of Molecular Biology*. doi:10.1006/jmbi.1999.3091

> A two-stage neural network has been used to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST. Despite the simplicity and convenience of the approach used, the results are found to be superior to those produced by other methods, including the popular PHD method according to our own benchmarking results and the results from the recent Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP3), where the method was evaluated by stringent blind testing. Using a new testing set based on a set of 187 unique folds, and three-way cross-validation based on structural similarity criteria rather than sequence similarity criteria used previously (no similar folds were present in both the testing and training sets) the method presented here (PSIPRED) achieved an average Q3 score of between 76.5% to 78.3% depending on the precise definition of observed secondary structure used, which is the highest published score for any method to date. Given the success of the method in CASP3, it is reasonable to be confident that the evaluation presented here gives a fair indication of the performance of the method in general.

- ⋆ Šali and Blundell (1993) — MODELLER.

Comparative Protein Modelling by Satisfaction of Spatial Restraints. Šali & Blundell (1993). *Journal of Molecular Biology*. doi:10.1006/jmbi.1993.1626

> We describe a comparative protein modelling method designed to find the most probable structure for a sequence given its alignment with related structures. The three-dimensional (3D) model is obtained by optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions (pdfs) for the features restrained. For example, the probabilities for main-chain conformations of a modelled residue may be restrained by its residue type, main-chain conformation of an equivalent residue in a related protein, and the local similarity between the two sequences. Several such pdfs are obtained from the correlations between structural features in 17 families of homologous proteins which have been aligned on the basis of their 3D structures. The pdfs restrain C alpha-C alpha distances, main-chain N-O distances, main-chain and side-chain dihedral angles. A smoothing procedure is used in the derivation of these relationships to minimize the problem of a sparse database. The 3D model of a protein is obtained by optimization of the molecular pdf such that the model violates the input restraints as little as possible. The molecular pdf is derived as a combination of pdfs restraining individual spatial features of the whole molecule. The optimization procedure is a variable target function method that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms. The method is automated and is illustrated by the modelling of trypsin from two other serine proteinases.

- Bowie et al. (1991), Jones et al. (1992), Fischer and Eisenberg (1996) — 3D profiles, threading and fold recognition.

A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. Bowie et al. (1991). *Science*. doi:10.1126/science.1853201

> The inverse protein folding problem, the problem of finding which amino acid sequences fold into a known three-dimensional (3D) structure, can be effectively attacked by finding sequences that are most compatible with the environments of the residues in the 3D structure. The environments are described by: (i) the area of the residue buried in the protein and inaccessible to solvent; (ii) the fraction of side-chain area that is covered by polar atoms (O and N); and (iii) the local secondary structure. Examples of this 3D profile method are presented for four families of proteins: the globins, cyclic AMP (adenosine 3',5'-monophosphate) receptor-like proteins, the periplasmic binding proteins, and the actins. This method is able to detect the structural similarity of the actins and 70- kilodalton heat shock proteins, even though these protein families share no detectable sequence similarity.

A new approach to protein fold recognition. Jones et al. (1992). *Nature*. doi:10.1038/358086a0

> The prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs or folds in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

Protein fold recognition using sequence-derived predictions. Fischer & Eisenberg (1996). *Protein Science*. doi:10.1002/pro.5560050516

> In protein fold recognition, one assigns a probe amino acid sequence of unknown structure to one of a library of target 3D structures. Correct assignment depends on effective scoring of the probe sequence for its compatibility with each of the target structures. Here we show that, in addition to the amino acid sequence of the probe, sequence-derived properties of the probe sequence (such as the predicted secondary structure) are useful in fold assignment. The additional measure of compatibility between probe and target is the level of agreement between the predicted secondary structure of the probe and the known secondary structure of the target fold. That is, we recommend a sequence-structure compatibility function that combines previously developed compatibility functions (such as the 3D-1D scores of Bowie et al. [1991] or sequence-sequence replacement tables) with the predicted secondary structure of the probe sequence. The effect on fold assignment of adding predicted secondary structure is evaluated here by using a benchmark set of proteins (Fischer et al., 1996a). The 3D structures of the probe sequences of the benchmark are actually known, but are ignored by our method. The results show that the inclusion of the predicted secondary structure improves fold assignment by about 25%. The results also show that, if the true secondary structure of the probe were known, correct fold assignment would increase by an additional 8-32%. We conclude that incorporating sequence-derived predictions significantly improves assignment of sequences to known 3D folds. Finally, we apply the new method to assign folds to sequences in the SWISSPROT database; six fold assignments are given that are not detectable by standard sequence-sequence comparison methods; for two of these, the fold is known from X-ray crystallography and the fold assignment is correct.

- Simons et al. (1997), Jones (2001), ⋆ Bradley et al. (2005) — fragment assembly with Rosetta and FRAGFOLD.

Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. Simons et al. (1997). *Journal of Molecular Biology*. doi:10.1006/jmbi.1997.0959

> We explore the ability of a simple simulated annealing procedure to assemble native-like structures from fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. Environment and residue pair specific contributions to the scoring functions appear as the first two terms in a series expansion for the residue probability distributions in the protein database; the decoupling of the distance and environment dependencies of the distributions resolves the major problems with current database-derived scoring functions noted by Thomas and Dill. The simulated annealing procedure rapidly and frequently generates native-like structures for small helical proteins and better than random structures for small beta sheet containing proteins. Most of the simulated structures have native-like solvent accessibility and secondary structure patterns, and thus ensembles of these structures provide a particularly challenging set of decoys for evaluating scoring functions. We investigate the effects of multiple sequence information and different types of conformational constraints on the overall performance of the method, and the ability of a variety of recently developed scoring functions to recognize the native-like conformations in the ensembles of simulated structures.

Predicting novel protein folds by using FRAGFOLD. Jones (2001). *Proteins: Structure, Function, and Genetics*. doi:10.1002/prot.1171

> The results of applying a fragment-based protein tertiary structure prediction method to the prediction of 8 CASP4 targets are described. The method is based on the assembly of supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. Despite the significant degree of success in this case, there is clearly much more developmental work required

before predictions with the accuracy of a good homology model, or even a good fold recognition model, can be made with use of this kind of approach.

Toward High-Resolution de Novo Structure Prediction for Small Proteins. Bradley et al. (2005). *Science*. doi:10.1126/science.1113801

> The prediction of protein structure from amino acid sequence is a grand challenge of computational molecular biology. By using a combination of improved low- and high-resolution conformational sampling methods, improved atomically detailed potential functions that capture the jigsaw puzzle-like packing of protein cores, and high-performance computing, high-resolution structure prediction (<1.5 angstroms) can be achieved for small protein domains (<85 residues). The primary bottleneck to consistent high-resolution prediction appears to be conformational sampling.

B.9.6 Coevolution and deep learning

- Göbel et al. (1994) — correlated mutations.

Correlated mutations and residue contacts in proteins. Göbel et al. (1994). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.340180402

> The maintenance of protein function and structure constrains the evolution of amino acid sequences. This fact can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions. Here we present a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. We then use these correlations to predict contact maps for each of 11 protein families and compare the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranges from 37 to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. Predicted contact maps can be used as input for the calculation of protein tertiary structure, either from sequence information alone or in combination with experimental information.

- Weigt et al. (2009), Morcos et al. (2011), Marks et al. (2011), Jones et al. (2011) — DCA and PSICOV.

Identification of direct residue contacts in protein–protein interaction by message passing. Weigt et al. (2009). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0805923106

> Understanding the molecular determinants of specificity in protein-protein interaction is an outstanding challenge of postgenome biology. The availability of large protein databases generated from sequences of hundreds of bacterial genomes enables various statistical approaches to this problem. In this context covariance-based methods have been used to identify correlation between amino acid positions in interacting proteins. However, these methods have an important shortcoming, in that they cannot distinguish between directly and indirectly correlated residues. We developed a method that combines covariance analysis with global inference analysis, adopted from use in statistical physics. Applied to a set of >2,500 representatives of the bacterial two-component signal transduction system, the combination of covariance with global inference successfully and robustly identified residue pairs that are proximal in space without resorting to ad hoc tuning parameters, both for heterointeractions between sensor kinase (SK) and response regulator (RR) proteins and for homointeractions between RR proteins. The spectacular success of this approach illustrates the effectiveness of the global inference approach in identifying direct interaction based on sequence information alone. We expect this method to be applicable soon to interaction surfaces between proteins present in only 1 copy per genome as the number of sequenced genomes continues to expand. Use of this method could significantly increase the potential targets for therapeutic intervention, shed light on the mechanism of protein-protein interaction, and establish the foundation for the accurate prediction of interacting protein partners.

Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Morcos et al. (2011). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1111471108

The similarity in the three-dimensional structures of homologous proteins imposes strong constraints on their sequence variability. It has long been suggested that the resulting correlations among amino acid compositions at different sequence positions can be exploited to infer spatial contacts within the tertiary protein structure. Crucial to this inference is the ability to disentangle direct and indirect correlations, as accomplished by the recently introduced direct-coupling analysis (DCA). Here we develop a computationally efficient implementation of DCA, which allows us to evaluate the accuracy of contact prediction by DCA for a large number of protein domains, based purely on sequence information. DCA is shown to yield a large number of correctly predicted contacts, recapitulating the global structure of the contact map for the majority of the protein domains examined. Furthermore, our analysis captures clear signals beyond intradomain residue contacts, arising, e.g., from alternative protein conformations, ligand-mediated residue couplings, and interdomain interactions in protein oligomers. Our findings suggest that contacts predicted by DCA can be used as a reliable guide to facilitate computational predictions of alternative protein conformations, protein complex formation, and even the de novo prediction of protein domain structures, contingent on the existence of a large number of homologous sequences which are being rapidly made available due to advances in genome sequencing.

Protein 3D Structure Computed from Evolutionary Sequence Variation. Marks et al. (2011). *PLoS ONE*. doi:10.1371/journal.pone.0028766

The evolutionary trajectory of a protein through sequence space is constrained by its function. Collections of sequence homologs record the outcomes of millions of evolutionary experiments in which the protein evolves according to these constraints. Deciphering the evolutionary record held in these sequences and exploiting it for predictive and engineering purposes presents a formidable challenge. The potential benefit of solving this challenge is amplified by the advent of inexpensive high-throughput genomic sequencing.In this paper we ask whether we can infer evolutionary constraints from a set of sequence homologs of a protein. The challenge is to distinguish true co-evolution couplings from the noisy set of observed correlations. We address this challenge using a maximum entropy model of the protein sequence, constrained by the statistics of the multiple sequence alignment, to infer residue pair couplings. Surprisingly, we find that the strength of these inferred couplings is an excellent predictor of residue-residue proximity in folded structures. Indeed, the top-scoring residue couplings are sufficiently accurate and well-distributed to define the 3D protein fold with remarkable accuracy.We quantify this observation by computing, from sequence alone, all-atom 3D structures of fifteen test proteins from different fold classes, ranging in size from 50 to 260 residues, including a G-protein coupled receptor. These blinded inferences are de novo, i.e., they do not use homology modeling or sequence-similar fragments from known structures. The co-evolution signals provide sufficient information to determine accurate 3D protein structure to 2.7-4.8 Å C(α)-RMSD error relative to the observed structure, over at least two-thirds of the protein (method called EVfold, details at http://EVfold.org). This discovery provides insight into essential interactions constraining protein evolution and will facilitate a comprehensive survey of the universe of protein structures, new strategies in protein and drug design, and the identification of functional genetic variants in normal and disease genomes.

PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Jones et al. (2011). *Bioinformatics*. doi:10.1093/bioinformatics/btr638

Motivation The accurate prediction of residue-residue contacts, critical for maintaining the native fold of a protein, remains an open problem in the field of structural bioinformatics. Interest in this long-standing problem has increased recently with algorithmic improvements and the rapid growth in the sizes of sequence families. Progress could have major impacts in both structure and function prediction to name but two benefits. Sequence-based contact predictions are usually made by identifying correlated mutations within multiple sequence alignments (MSAs), most commonly through the information-theoretic approach of calculating mutual information between pairs of sites in proteins. These predictions are often inaccurate because the true covariation signal in the MSA is often masked by biases from many ancillary indirect-coupling or phylogenetic effects. Here we present a novel method, PSICOV, which introduces the use of sparse inverse covariance estimation to the problem of protein contact prediction. Our method builds on work which had previously demonstrated corrections for phylogenetic and entropic correlation noise and allows accurate discrimination of direct from indirectly coupled mutation correlations in the MSA. Results PSICOV displays a mean precision substantially better than the best performing normalized mutual information approach and Bayesian networks. For 118 out of 150 targets, the L/5 (i.e. top-L/5 predictions for a protein of length L) precision for long-range contacts (sequence separation >23) was ≥ 0.5, which represents an improvement sufficient to be of significant benefit in protein structure prediction or model quality assessment. Availability The PSICOV source code can be downloaded from http://bioinf.cs.ucl.ac.uk/downloads/PSICOV.

- Skwark et al. (2014), ⋆ Wang et al. (2017) — contact prediction with machine learning.

Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. Skwark et al. (2014). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1003889

> Given sufficient large protein families, and using a global statistical inference approach, it is possible to obtain sufficient accuracy in protein residue contact predictions to predict the structure of many proteins. However, these approaches do not consider the fact that the contacts in a protein are neither randomly, nor independently distributed, but actually follow precise rules governed by the structure of the protein and thus are interdependent. Here, we present PconsC2, a novel method that uses a deep learning approach to identify protein-like contact patterns to improve contact predictions. A substantial enhancement can be seen for all contacts independently on the number of aligned sequences, residue separation or secondary structure type, but is largest for β-sheet containing proteins. In addition to being superior to earlier methods based on statistical inferences, in comparison to state of the art methods using machine learning, PconsC2 is superior for families with more than 100 effective sequence homologs. The improved contact prediction enables improved structure prediction.

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. Wang et al. (2017). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1005324

> Motivation Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still of low quality and not very useful for de novo structure prediction. Method This paper presents a new deep learning method that predicts contacts by integrating both evolutionary coupling (EC) and sequence conservation information through an ultra-deep neural network formed by two deep residual neural networks. The first residual network conducts a series of 1-dimensional convolutional transformation of sequential features; the second residual network conducts a series of 2-dimensional convolutional transformation of pairwise information including output of the first residual network, EC information and pairwise potential. By using very deep residual networks, we can accurately model contact occurrence patterns and complex sequence-structure relationship and thus, obtain higher-quality contact prediction regardless of how many sequence homologs are available for proteins in question. Results Our method greatly outperforms existing methods and leads to much more accurate contact-assisted folding. Tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398 membrane proteins, the average top L long-range prediction accuracy obtained by our method, one representative EC method CCMpred and the CASP11 winner MetaPSICOV is 0.47, 0.21 and 0.30, respectively; the average top L/10 long-range accuracy of our method, CCMpred and MetaPSICOV is 0.77, 0.47 and 0.59, respectively. Ab initio folding using our predicted contacts as restraints but without any force fields can yield correct folds (i.e., TMscore>0.6) for 203 of the 579 test proteins, while that using MetaPSICOV- and CCMpred-predicted contacts can do so for only 79 and 62 of them, respectively. Our contact-assisted models also have much better quality than template-based models especially for membrane proteins. The 3D models built from our contact prediction have TMscore>0.5 for 208 of the 398 membrane proteins, while those from homology modeling have TMscore>0.5 for only 10 of them. Further, even if trained mostly by soluble proteins, our deep learning method works very well on membrane proteins. In the recent blind CAMEO benchmark, our fully-automated web server implementing this method successfully folded 6 targets with a new fold and only 0.3L-2.3L effective sequence homologs, including one β protein of 182 residues, one α+β protein of 125 residues, one α protein of 140 residues, one α protein of 217 residues, one α/β of 260 residues and one α protein of 462 residues. Our method also achieved the highest F1 score on free-modeling targets in the latest CASP (Critical Assessment of Structure Prediction), although it was not fully implemented back then. Availability http://raptorx.uchicago.edu/ContactMap/.

- Ovchinnikov et al. (2017) — structures from metagenome sequences.

Protein structure determination using metagenome sequence data. Ovchinnikov et al. (2017). *Science*. doi:10.1126/science.aah4043

> Despite decades of work by structural biologists, there are still ~5200 protein families with unknown structure outside the range of comparative modeling. We show that Rosetta structure prediction guided by residue-residue contacts inferred from evolutionary information can accurately model proteins that belong to large families and that metagenome sequence data more than triple the number of protein families with sufficient sequences for accurate modeling. We then integrate metagenome data, contact-based structure matching, and Rosetta structure calculations to generate models for 614 protein families with currently unknown structures; 206 are membrane proteins and 137 have folds not represented in the Protein Data

> Bank. This approach provides the representative models for large protein families originally envisioned as the goal of the Protein Structure Initiative at a fraction of the cost.

- Senior et al. (2020) — AlphaFold at CASP13.

  Improved protein structure prediction using potentials from deep learning. Senior et al. (2020). *Nature*. doi:10.1038/s41586-019-1923-7

  > Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence 1 . This problem is of fundamental importance as the structure of a protein largely determines its function 2 ; however, protein structures can be difficult to determine experimentally. Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures 3 . Here we show that we can train a neural network to make accurate predictions of the distances between pairs of residues, which convey more information about the structure than contact predictions. Using this information, we construct a potential of mean force 4 that can accurately describe the shape of a protein. We find that the resulting potential can be optimized by a simple gradient descent algorithm to generate structures without complex sampling procedures. The resulting system, named AlphaFold, achieves high accuracy, even for sequences with fewer homologous sequences. In the recent Critical Assessment of Protein Structure Prediction 5 (CASP13)-a blind assessment of the state of the field-AlphaFold created high-accuracy structures (with template modelling (TM) scores 6 of 0.7 or higher) for 24 out of 43 free modelling domains, whereas the next best method, which used sampling and contact information, achieved such accuracy for only 14 out of 43 domains. AlphaFold represents a considerable advance in protein-structure prediction. We expect this increased accuracy to enable insights into the function and malfunction of proteins, especially in cases for which no structures for homologous proteins have been experimentally determined 7 .

- Baek et al. (2021) — RoseTTAFold.

  Accurate prediction of protein structures and interactions using a three-track neural network. Baek et al. (2021). *Science*. doi:10.1126/science.abj8754

  > DeepMind presented notably accurate predictions at the recent 14th Critical Assessment of Structure Prediction (CASP14) conference. We explored network architectures that incorporate related ideas and obtained the best performance with a three-track network in which information at the one-dimensional (1D) sequence level, the 2D distance map level, and the 3D coordinate level is successively transformed and integrated. The three-track network produces structure predictions with accuracies approaching those of DeepMind in CASP14, enables the rapid solution of challenging x-ray crystallography and cryo-electron microscopy structure modeling problems, and provides insights into the functions of proteins of currently unknown structure. The network also enables rapid generation of accurate protein-protein complex models from sequence information alone, short-circuiting traditional approaches that require modeling of individual subunits followed by docking. We make the method available to the scientific community to speed biological research.

- Mirdita et al. (2022), Ahdritz et al. (2024) — ColabFold and OpenFold.

  ColabFold: making protein folding accessible to all. Mirdita et al. (2022). *Nature Methods*. doi:10.1038/s41592-022-01488-1

  > ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold. ColabFold's 40-60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit. Coupled with Google Colaboratory, ColabFold becomes a free and accessible platform for protein folding. ColabFold is open-source software available at https://github.com/sokrypton/ColabFold and its novel environmental databases are available at https://colabfold.mmseqs.com .

  OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Ahdritz et al. (2024). *Nature Methods*. doi:10.1038/s41592-024-02272-z

> AlphaFold2 revolutionized structural biology with the ability to predict protein structures with exceptionally high accuracy. Its implementation, however, lacks the code and data required to train new models. These are necessary to (1) tackle new tasks, like protein-ligand complex structure prediction, (2) investigate the process by which the model learns and (3) assess the model's capacity to generalize to unseen regions of fold space. Here we report OpenFold, a fast, memory efficient and trainable implementation of AlphaFold2. We train OpenFold from scratch, matching the accuracy of AlphaFold2. Having established parity, we find that OpenFold is remarkably robust at generalizing even when the size and diversity of its training set is deliberately limited, including near-complete elisions of classes of secondary structure elements. By analyzing intermediate structures produced during training, we also gain insights into the hierarchical manner in which OpenFold learns to fold. In sum, our studies demonstrate the power and utility of OpenFold, which we believe will prove to be a crucial resource for the protein modeling community.

- Evans et al. (2021), Bryant, Pozzati, and Elofsson (2022) — AlphaFold-Multimer and complex prediction.

  Protein complex prediction with AlphaFold-Multimer. Evans et al. (2021). *bioRxiv (preprint)*. doi:10.1101/2021.10.04.463034

  > While the vast majority of well-structured single protein chains can now be predicted to high accuracy due to the recent AlphaFold [1] model, the prediction of multi-chain protein complexes remains a challenge in many cases. In this work, we demonstrate that an AlphaFold model trained specifically for multimeric inputs of known stoichiometry, which we call AlphaFold-Multimer, significantly increases accuracy of predicted multimeric interfaces over input-adapted single-chain AlphaFold while maintaining high intra-chain accuracy. On a benchmark dataset of 17 heterodimer proteins without templates (introduced in [2]) we achieve at least medium accuracy (DockQ [3] ≥ 0.49) on 13 targets and high accuracy (DockQ ≥ 0.8) on 7 targets, compared to 9 targets of at least medium accuracy and 4 of high accuracy for the previous state of the art system (an AlphaFold-based system from [2]). We also predict structures for a large dataset of 4,446 recent protein complexes, from which we score all non-redundant interfaces with low template identity. For heteromeric interfaces we successfully predict the interface (DockQ ≥ 0.23) in 70% of cases, and produce high accuracy predictions (DockQ ≥ 0.8) in 26% of cases, an improvement of +27 and +14 percentage points over the flexible linker modification of AlphaFold [4] respectively. For homomeric inter-faces we successfully predict the interface in 72% of cases, and produce high accuracy predictions in 36% of cases, an improvement of +8 and +7 percentage points respectively.

  Improved prediction of protein-protein interactions using AlphaFold2. Bryant et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-28865-w

  > Predicting the structure of interacting protein chains is a fundamental step towards understanding protein function. Unfortunately, no computational method can produce accurate structures of protein complexes. AlphaFold2, has shown unprecedented levels of accuracy in modelling single chain protein structures. Here, we apply AlphaFold2 for the prediction of heterodimeric protein complexes. We find that the AlphaFold2 protocol together with optimised multiple sequence alignments, generate models with acceptable quality (DockQ ≥ 0.23) for 63% of the dimers. From the predicted interfaces we create a simple function to predict the DockQ score which distinguishes acceptable from incorrect models as well as interacting from non-interacting proteins with state-of-art accuracy. We find that, using the predicted DockQ scores, we can identify 51% of all interacting pairs at 1% FPR.

- Boitreaud et al. (2024), Wohlwend et al. (2024), Passaro et al. (2025), Chen et al. (2025) — open AlphaFold3-like models (first released as preprints).

  Chai-1: Decoding the molecular interactions of life. Boitreaud et al. (2024). *bioRxiv (preprint)*. doi:10.1101/2024.10.10.615955

  > We introduce Chai-1, a multi-modal foundation model for molecular structure prediction that performs at the state-of-the-art across a variety of tasks relevant to drug discovery. Chai-1 can optionally be prompted with experimental restraints (e.g. derived from wet-lab data) which boosts performance by double-digit percentage points. Chai-1 can also be run in single-sequence mode with-out MSAs while preserving most of its performance. We release Chai-1 model weights and inference code as a Python package for non-commercial use and via a web interface where it can be used for free including for commercial drug discovery purposes.

Boltz-1 Democratizing Biomolecular Interaction Modeling. Wohlwend et al. (2024). *bioRxiv (preprint)*. doi:10.1101/2024.11.19.624167

> Understanding biomolecular interactions is fundamental to advancing fields like drug discovery and protein design. In this paper, we introduce B oltz -1, an open-source deep learning model incorporating innovations in model architecture, speed optimization, and data processing achieving A lpha F old 3-level accuracy in predicting the 3D structures of biomolecular complexes. B oltz -1 demonstrates a performance on-par with state-of-the-art commercial models on a range of diverse benchmarks, setting a new benchmark for commercially accessible tools in structural biology. Further, we push the boundary of capabilities of these models with B oltz-steering , a new inference time steering technique that is able to fix hallucinations and non-physical predictions from the models. By releasing the training and inference code, model weights, datasets, and benchmarks under the MIT open license, we aim to foster global collaboration, accelerate discoveries, and provide a robust platform for advancing biomolecular modeling.

Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. Passaro et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.06.14.659707

> Accurately modeling biomolecular interactions is a central challenge in modern biology. While recent advances, such as AlphaFold3 and Boltz-1, have substantially improved our ability to predict biomolecular complex structures, these models still fall short in predicting binding affinity, a critical property underlying molecular function and therapeutic efficacy. Here, we present Boltz-2, a new structural biology foundation model that exhibits strong performance for both structure and affinity prediction. Boltz-2 introduces controllability features including experimental method conditioning, distance constraints, and multi-chain template integration for structure prediction, and is, to our knowledge, the first AI model to approach the performance of free-energy perturbation (FEP) methods in estimating small molecule–protein binding affinity. Crucially, it achieves strong correlation with experimental readouts on many benchmarks, while being at least 1000 × more computationally efficient than FEP. By coupling Boltz-2 with a generative model for small molecules, we demonstrate an effective workflow to find diverse, synthesizable, high-affinity binders, as estimated by absolute FEP simulations on the TYK2 target. To foster broad adoption and further innovation at the intersection of machine learning and biology, we are releasing Boltz-2 weights, inference, and training code 1 under a permissive open license, providing a robust and extensible foundation for both academic and industrial research.

Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. Chen et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.01.08.631967

> In this technical report, we present Protenix, a comprehensive reproduction of AlphaFold3 (AF3), aimed at advancing the field of biomolecular structure prediction. Protenix tackles the challenges of predicting complex interactions involving proteins, ligands, and nucleic acids, while enhancing accessibility and reproducibility. Across diverse benchmarks, including PoseBusters V2, low-homology PDB sets, and CASP15 RNA, Protenix achieves state-of-the-art performance in protein-ligand, protein-protein, and protein-nucleic acid predictions. We also address limitations, such as potential memorization effects, and outline future directions for improvement. By open-sourcing Protenix, we aim to democratize advanced structure prediction tools and accelerate interdisciplinary research in computational biology and drug discovery.

- Isomorphic Labs Team (2026), OpenDDE contributors (2026) — the IsoDDE technical report (February 2026, not peer reviewed) and the OpenDDE software pipeline (GitHub, 2026).

  Accurate Predictions of Novel Biomolecular Interactions with IsoDDE. Isomorphic Labs Team (2026). *Isomorphic Labs*. doi:10.5281/zenodo.19699685

  OpenDDE: Open Drug Discovery Engine. OpenDDE contributors (2026). *GitHub repository*. https://github.com/contact-ajmal/opendde

- Wayment-Steele et al. (2023) — predicting multiple conformations.

  Predicting multiple conformations via sequence clustering and AlphaFold2. Wayment-Steele et al. (2023). *Nature*. doi:10.1038/s41586-023-06832-9

AlphaFold2 (ref. 1) has revolutionized structural biology by accurately predicting single structures of proteins. However, a protein's biological function often depends on multiple conformational substates 2 , and disease-causing point mutations often cause population changes within these substates 3,4 . We demonstrate that clustering a multiple-sequence alignment by sequence similarity enables AlphaFold2 to sample alternative states of known metamorphic proteins with high confidence. Using this method, named AF-Cluster, we investigated the evolutionary distribution of predicted structures for the metamorphic protein KaiB 5 and found that predictions of both conformations were distributed in clusters across the KaiB family. We used nuclear magnetic resonance spectroscopy to confirm an AF-Cluster prediction: a cyanobacteria KaiB variant is stabilized in the opposite state compared with the more widely studied variant. To test AF-Cluster's sensitivity to point mutations, we designed and experimentally verified a set of three mutations predicted to flip KaiB from Rhodobacter sphaeroides from the ground to the fold-switched state. Finally, screening for alternative states in protein families without known fold switching identified a putative alternative state for the oxidoreductase Mpt53 in Mycobacterium tuberculosis. Further development of such bioinformatic methods in tandem with experiments will probably have a considerable impact on predicting protein energy landscapes, essential for illuminating biological function.

B.9.7 Language models

- EvolutionaryScale (2024) — ESM C.

  ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale (2024). https://www.evolutionaryscale.ai/blog/esm-cambrian

- Candido et al. (2026) — ESMFold2 (bioRxiv preprint, June 2026).

  Language Modeling Materializes a World Model of Protein Biology. Candido et al. (2026). *bioRxiv (preprint)*. doi:10.64898/2026.06.03.729735

  > Proteins are fundamental to life. The full extent of their biology is beyond our ability to characterize with experimental approaches in the physical laboratory. Accurate digital representations could accelerate the discovery of protein biology through virtual experiments. We propose language modeling to learn unified and general representations that can be scaled to all of protein biology. Building on these representations, we develop a structure prediction model that exceeds the performance of established methods for biomolecular complex prediction across benchmarks, including for the interactions of antibodies with their targets. A simple search procedure yields high experimental success rates for the discovery of proteins with nanomolar binding affinities for both miniproteins and single-chain antibodies, a modality critical for therapeutic design. Study of the concepts in the language model's representation space reveals a systematic organization aligned with the reductionist understanding of proteins developed through empirical science. Leveraging this organization, we generate a comprehensive map of protein biology encompassing over 6.8 billion sequences and 1.1 billion predicted structures, identifying connections across known and unknown biology. As a whole, this shows language modeling as a powerful substrate for representing the biology of proteins, operating across scales from the prediction and design of protein interactions at the atomic level, to identifying properties of proteins at different levels of granularity and abstraction, to the scale of mapping connections between proteins across billions of years of evolution.

- Hayes et al. (2025) — ESM3.

  Simulating 500 million years of evolution with a language model. Hayes et al. (2025). *Science*. doi:10.1126/science.ads0018

  > More than 3 billion years of evolution have produced an image of biology encoded into the space of natural proteins. Here, we show that language models trained at scale on evolutionary data can generate functional proteins that are far away from known proteins. We present ESM3, a frontier multimodal generative language model that reasons over the sequence, structure, and function of proteins. ESM3 can follow complex prompts combining its modalities and is highly responsive to alignment to improve its fidelity. We have prompted ESM3 to generate fluorescent proteins. Among the generations that we synthesized, we found a bright fluorescent protein at a far distance (58% sequence identity) from known fluorescent proteins, which we estimate is equivalent to simulating 500 million years of evolution.

- Heinzinger et al. (2024) — ProstT5.

  Bilingual language model for protein sequence and structure. Heinzinger et al. (2024). *NAR Genomics and Bioinformatics*. doi:10.1093/nargab/lqae150

  > Adapting language models to protein sequences spawned the development of powerful protein language models (pLMs). Concurrently, AlphaFold2 broke through in protein structure prediction. Now we can systematically and comprehensively explore the dual nature of proteins that act and exist as three-dimensional (3D) machines and evolve as linear strings of one-dimensional (1D) sequences. Here, we leverage pLMs to simultaneously model both modalities in a single model. We encode protein structures as token sequences using the 3Di-alphabet introduced by the 3D-alignment method Foldseek . For training, we built a non-redundant dataset from AlphaFoldDB and fine-tuned an existing pLM (ProtT5) to translate between 3Di and amino acid sequences. As a proof-of-concept for our novel approach, dubbed Protein 'structure-sequence' T5 (ProstT5), we showed improved performance for subsequent, structure-related prediction tasks, leading to three orders of magnitude speedup for deriving 3Di. This will be crucial for future applications trying to search metagenomic sequence databases at the sensitivity of structure comparisons. Our work showcased the potential of pLMs to tap into the information-rich protein structure revolution fueled by AlphaFold2. ProstT5 paves the way to develop new tools integrating the vast resource of 3D predictions and opens new research avenues in the post-AlphaFold2 era.

B.9.8 Membrane proteins

- ⋆ Heijne (1992) — hydrophobicity analysis and the positive-inside rule.

  Membrane protein structure prediction. von Heijne (1992). *Journal of Molecular Biology*. doi:10.1016/0022-2836(92)90934-c

  > A new strategy for predicting the topology of bacterial inner membrane proteins is proposed on the basis of hydrophobicity analysis, automatic generation of a set of possible topologies and ranking of these according to the positive-inside rule. A straightforward implementation with no attempts at optimization predicts the correct topology for 23 out of 24 inner membrane proteins with experimentally determined topologies, and correctly identifies 135 transmembrane segments with only one overprediction.

- Krogh et al. (2001) — TMHMM.

  Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Krogh et al. (2001). *Journal of Molecular Biology*. doi:10.1006/jmbi.2000.4315

  > We describe and validate a new membrane protein topology prediction method, TMHMM, based on a hidden Markov model. We present a detailed analysis of TMHMM's performance, and show that it correctly predicts 97-98 % of the transmembrane helices. Additionally, TMHMM can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99 %, although the accuracy drops when signal peptides are present. This high degree of accuracy allowed us to predict reliably integral membrane proteins in a large collection of genomes. Based on these predictions, we estimate that 20-30 % of all genes in most genomes encode membrane proteins, which is in agreement with previous estimates. We further discovered that proteins with N(in)-C(in) topologies are strongly preferred in all examined organisms, except Caenorhabditis elegans, where the large number of 7TM receptors increases the counts for N(out)-C(in) topologies. We discuss the possible relevance of this finding for our understanding of membrane protein assembly mechanisms. A TMHMM prediction service is available at http://www.cbs.dtu.dk/services/TMHMM/.

- Elofsson and Heijne (2007) — membrane protein structure: prediction versus reality.

  Membrane Protein Structure: Prediction versus Reality. Elofsson & Heijne (2007). *Annual Review of Biochemistry*. doi:10.1146/annurev.biochem.76.052705.163539

  > Since high-resolution structural data are still scarce, different kinds of theoretical structure prediction algorithms are of major importance in membrane protein biochemistry. But how well do the current prediction methods perform? Which structural features can be predicted and which cannot? And what can we expect in the next few years?

- Hallgren et al. (2022) — DeepTMHMM.

  DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. Hallgren et al. (2022). *bioRxiv (preprint)*. doi:10.1101/2022.04.08.487609

  > Transmembrane proteins span the lipid bilayer and are divided into two major structural classes, namely alpha helical and beta barrels. We introduce DeepTMHMM, a deep learning protein language model-based algorithm that can detect and predict the topology of both alpha helical and beta barrels proteins with unprecedented accuracy. DeepTMHMM (https://dtu.biolib.com/DeepTMHMM) scales to proteomes and covers all domains of life, which makes it ideal for metagenomics analyses.

- Wimley (2003), Bagos et al. (2004), Hayat et al. (2016) — β-barrel membrane proteins and their prediction.

  The versatile β-barrel membrane protein. Wimley (2003). *Current Opinion in Structural Biology*. doi:10.1016/s0959-440x(03)00099-x

  > The beta-barrel membrane protein is found in the outer membranes of bacteria, mitochondria and chloroplasts. Approximately 2-3% of the genes in Gram-negative bacterial genomes encode beta-barrels. Whereas there are fewer than 20 known three-dimensional beta-barrel structures, genomic databases currently contain thousands of beta-barrels belonging to dozens of families. New research is revealing the variety of beta-barrel structures and the variety of functions performed by these versatile proteins.

  PRED-TMBB: a web server for predicting the topology of -barrel outer membrane proteins. Bagos et al. (2004). *Nucleic Acids Research*. doi:10.1093/nar/gkh417

  > The beta-barrel outer membrane proteins constitute one of the two known structural classes of membrane proteins. Whereas there are several different web-based predictors for alpha-helical membrane proteins, currently there is no freely available prediction method for beta-barrel membrane proteins, at least with an acceptable level of accuracy. We present here a web server (PRED-TMBB, http://bioinformatics.biol.uoa.gr/PRED-TMBB) which is capable of predicting the transmembrane strands and the topology of beta-barrel outer membrane proteins of Gram-negative bacteria. The method is based on a Hidden Markov Model, trained according to the Conditional Maximum Likelihood criterion. The model was retrained and the training set now includes 16 non-homologous outer membrane proteins with structures known at atomic resolution. The user may submit one sequence at a time and has the option of choosing between three different decoding methods. The server reports the predicted topology of a given protein, a score indicating the probability of the protein being an outer membrane beta-barrel protein, posterior probabilities for the transmembrane strand prediction and a graphical representation of the assumed position of the transmembrane strands with respect to the lipid bilayer.

  Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. Hayat et al. (2016). *Bioinformatics*. doi:10.1093/bioinformatics/btw025

  > Unlabelled : Accurate topology prediction of transmembrane β-barrels is still an open question. Here, we present BOCTOPUS2, an improved topology prediction method for transmembrane β-barrels that can also identify the barrel domain, predict the topology and identify the orientation of residues in transmembrane β-strands. The major novelty of BOCTOPUS2 is the use of the dyad-repeat pattern of lipid and pore facing residues observed in transmembrane β-barrels. In a cross-validation test on a benchmark set of 42 proteins, BOCTOPUS2 predicts the correct topology in 69% of the proteins, an improvement of more than 10% over the best earlier method (BOCTOPUS) and in addition, it produces significantly fewer erroneous predictions on non-transmembrane β-barrel proteins. Availability and implementation BOCTOPUS2 webserver along with full dataset and source code is available at http://boctopus.bioinfo.se/ Contact : arne@bioinfo.se Supplementary information Supplementary data are available at Bioinformatics online.

B.9.9 Dynamics and docking

- Levitt and Warshel (1975), McCammon et al. (1977), Warshel and Levitt (1976) — the first simulations and QM/MM.

  Computer simulation of protein folding. Levitt & Warshel (1975). *Nature*. doi:10.1038/253694a0

> A new and very simple representation of protein conformations has been used together with energy minimisation and thermalisation to simulate protein folding. Under certain conditions, the method succeeds in ”renaturing” bovine pancreatic trypsin inhibitor from an open-chain conformation into a folded conformation close to that of the native molecule.

Dynamics of folded proteins. McCammon et al. (1977). *Nature*. doi:10.1038/267585a0

> The dynamics of a folded globular protein (bovine pancreatic trypsin inhibitor) have been studied by solving the equations of motion for the atoms with an empirical potential energy function. The results provide the magnitude, correlations and decay of fluctuations about the average structure. These suggest that the protein interior is fluid-like in that the local atom motions have a diffusional character.

Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. Warshel & Levitt (1976). *Journal of Molecular Biology*. doi:10.1016/0022-2836(76)90311-9

- Karplus and McCammon (2002) — molecular dynamics of biomolecules.

Molecular dynamics simulations of biomolecules. Karplus & McCammon (2002). *Nature Structural Biology*. doi:10.1038/nsb0902-646

> Molecular dynamics simulations are important tools for understanding the physical basis of the structure and function of biological macromolecules. The early view of proteins as relatively rigid structures has been replaced by a dynamic model in which the internal motions and resulting conformational changes play an essential role in their function. This review presents a brief description of the origin and early uses of biomolecular simulations. It then outlines some recent studies that illustrate the utility of such simulations and closes with a discussion of their ever-increasing potential for contributing to biology.

- Lindorff-Larsen et al. (2012) — validating force fields.

Systematic Validation of Protein Force Fields against Experimental Data. Lindorff-Larsen et al. (2012). *PLoS ONE*. doi:10.1371/journal.pone.0032131

> Molecular dynamics simulations provide a vehicle for capturing the structures, motions, and interactions of biological macromolecules in full atomic detail. The accuracy of such simulations, however, is critically dependent on the force field--the mathematical model used to approximate the atomic-level forces acting on the simulated molecular system. Here we present a systematic and extensive evaluation of eight different protein force fields based on comparisons of experimental data with molecular dynamics simulations that reach a previously inaccessible timescale. First, through extensive comparisons with experimental NMR data, we examined the force fields' abilities to describe the structure and fluctuations of folded proteins. Second, we quantified potential biases towards different secondary structure types by comparing experimental and simulation data for small peptides that preferentially populate either helical or sheet-like structures. Third, we tested the force fields' abilities to fold two small proteins--one α-helical, the other with β-sheet structure. The results suggest that force fields have improved over time, and that the most recent versions, while not perfect, provide an accurate description of many structural and dynamical properties of proteins.

- Tirion (1996), Bahar et al. (1997), Atilgan et al. (2001) — normal modes and elastic network models.

Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Tirion (1996). *Physical Review Letters*. doi:10.1103/physrevlett.77.1905

> Normal mode analysis (NMA) is a leading method for studying long-time dynamics and elasticity of biomolecules. The method proceeds from complex semiempirical potentials characterizing the covalent and noncovalent interactions between atoms. It is widely accepted that such detailed potentials are essential to the success of NMA's. We show that a single-parameter potential is sufficient to reproduce the slow dynamics in good detail. Costly and inaccurate energy minimizations are eliminated, permitting direct analysis of crystal coordinates. The technique can be used for new applications, such as mapping of one crystal form to another by means of slow modes, and studying anomalous dynamics of large proteins and complexes.

Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Bahar et al. (1997). *Folding and Design*. doi:10.1016/s1359-0278(97)00024-2

> Background An elastic network model is proposed for the interactions between closely (Results The temperature factors of the C alpha atoms of 12 X-ray structures, ranging from a 41 residue subunit to a 633 residue dimer, are accurately predicted. Cross-correlations are also efficiently characterized, in close agreement with results obtained with a normal mode analysis coupled with energy minimization. Conclusions The simple model and method proposed here provide a satisfactory description of the correlations between atomic fluctuations. Furthermore, this is achieved within computation times at least one order of magnitude shorter than commonly used molecular approaches.

Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. Atilgan et al. (2001). *Biophysical Journal*. doi:10.1016/s0006-3495(01)76033-x

> Fluctuations about the native conformation of proteins have proven to be suitably reproduced with a simple elastic network model, which has shown excellent agreement with a number of different properties for a wide variety of proteins. This scalar model simply investigates the magnitudes of motion of individual residues in the structure. To use the elastic model approach further for developing the details of protein mechanisms, it becomes essential to expand this model to include the added details of the directions of individual residue fluctuations. In this paper a new tool is presented for this purpose and applied to the retinol-binding protein, which indicates enhanced flexibility in the region of entry to the ligand binding site and for the portion of the protein binding to its carrier protein.

- Kuntz et al. (1982), Meng et al. (1992) — DOCK.

A geometric approach to macromolecule-ligand interactions. Kuntz et al. (1982). *Journal of Molecular Biology*. doi:10.1016/0022-2836(82)90153-x

Automated docking with grid-based energy evaluation. Meng et al. (1992). *Journal of Computational Chemistry*. doi:10.1002/jcc.540130412

> The ability to generate feasible binding orientations of a small molecule within a site of known structure is important for ligand design. We present a method that combines a rapid, geometric docking algorithm with the evaluation of molecular mechanics interaction energies. The computational costs of evaluation are minimal because we precalculate the receptor-dependent terms in the potential function at points on a three-dimensional grid. In four test cases where the components of crystallographically determined complexes are redocked, the "force field" score correctly identifies the family of orientations closest to the experimental binding geometry. Scoring functions that consider only steric factors or only electrostatic factors are less successful. The force field function will play an important role in our efforts to search databases for potential lead compounds.

- Böhm (1994), ⋆ Muegge and Martin (1999) — empirical and knowledge-based scoring functions.

The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. Böhm (1994). *Journal of Computer-Aided Molecular Design*. doi:10.1007/bf00126743

> A new simple empirical function has been developed that estimates the free energy of binding for a given protein-ligand complex of known 3D structure. The function takes into account hydrogen bonds, ionic interactions, the lipophilic protein-ligand contact surface and the number of rotatable bonds in the ligand. The dataset for the calibration of the function consists of 45 protein-ligand complexes. The new energy function reproduces the binding constants (ranging from 2.5.10(-2) to 4.10(-14) M, corresponding to binding energies between -9 and -76 kJ/mol) of the dataset with a standard deviation of 7.9 kJ/mol, corresponding to 1.4 orders of magnitude in binding affinity. The individual contributions to protein-ligand binding obtained from the scoring function are: ideal neutral hydrogen bond: -4.7 kJ/mol; ideal ionic interaction: -8.3 kJ/mol; lipophilic contact: -0.17 kJ/mol A2; one rotatable bond in the ligand: +1.4 kJ/mol. The function also contains a constant contribution (+5.4 kJ/mol) which may be rationalized as loss of translational and rotational entropy. The function can be evaluated very fast and is therefore also suitable for application in a 3D database search or de novo ligand design program such as LUDI.

A General and Fast Scoring Function for Protein–Ligand Interactions: A Simplified Potential Approach. Muegge & Martin (1999). *Journal of Medicinal Chemistry*. doi:10.1021/jm980536j

> A fast, simplified potential-based approach is presented that estimates the protein-ligand binding affinity based on the given 3D structure of a protein-ligand complex. This general, knowledge-based approach exploits structural information of known protein-ligand complexes extracted from the Brookhaven Protein Data Bank and converts it into distance-dependent Helmholtz free interaction energies of protein-ligand atom pairs (potentials of mean force, PMF). The definition of an appropriate reference state and the introduction of a correction term accounting for the volume taken by the ligand were found to be crucial for deriving the relevant interaction potentials that treat solvation and entropic contributions implicitly. A significant correlation between experimental binding affinities and computed score was found for sets of diverse protein-ligand complexes and for sets of different ligands bound to the same target. For 77 protein-ligand complexes taken from the Brookhaven Protein Data Bank, the calculated score showed a standard deviation from observed binding affinities of 1.8 log Ki units and an R2 value of 0.61. The best results were obtained for the subset of 16 serine protease complexes with a standard deviation of 1.0 log Ki unit and an R2 value of 0.86. A set of 33 inhibitors modeled into a crystal structure of HIV-1 protease yielded a standard deviation of 0.8 log Ki units from measured inhibition constants and an R2 value of 0.74. In contrast to empirical scoring functions that show similar or sometimes better correlation with observed binding affinities, our method does not involve deriving specific parameters that fit the observed binding affinities of protein-ligand complexes of a given training set. We compared the performance of the PMF score, Böhm's score (LUDI), and the SMOG score for eight different test sets of protein-ligand complexes. It was found that for the majority of test sets the PMF score performs best. The strength of the new approach presented here lies in its generality as no knowledge about measured binding affinities is needed to derive atomic interaction potentials. The use of the new scoring function in docking studies is outlined.

- Warren et al. (2005) — a critical assessment of docking.

  A Critical Assessment of Docking Programs and Scoring Functions. Warren et al. (2005). *Journal of Medicinal Chemistry*. doi:10.1021/jm050362n

  > Docking is a computational technique that samples conformations of small molecules in protein binding sites; scoring functions are used to assess which of these conformations best complements the protein binding site. An evaluation of 10 docking programs and 37 scoring functions was conducted against eight proteins of seven protein types for three tasks: binding mode prediction, virtual screening for lead identification, and rank-ordering by affinity for lead optimization. All of the docking programs were able to generate ligand conformations similar to crystallographically determined protein/ligand complex structures for at least one of the targets. However, scoring functions were less successful at distinguishing the crystallographic conformation from the set of docked poses. Docking programs identified active compounds from a pharmaceutically relevant pool of decoy compounds; however, no single program performed well for all of the targets. For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity.

- Trott and Olson (2009) — AutoDock Vina.

  AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Trott & Olson (2009). *Journal of Computational Chemistry*. doi:10.1002/jcc.21334

  > AutoDock Vina, a new program for molecular docking and virtual screening, is presented. AutoDock Vina achieves an approximately two orders of magnitude speed-up compared with the molecular docking software previously developed in our lab (AutoDock 4), while also significantly improving the accuracy of the binding mode predictions, judging by our tests on the training set used in AutoDock 4 development. Further speed-up is achieved from parallelism, by using multithreading on multicore machines. AutoDock Vina automatically calculates the grid maps and clusters the results in a way transparent to the user.

- Katchalski-Katzir et al. (1992) — FFT-based protein–protein docking.

  Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.. Katchalski-Katzir et al. (1992). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.89.6.2195

  > A geometric recognition algorithm was developed to identify molecular surface complementarity. It is based on a purely geometric approach and takes advantage of techniques applied in the field of pattern recognition. The algorithm involves an automated procedure including (i) a digital representation of the molecules (derived from atomic coordinates) by three-dimensional discrete functions that distinguishes

between the surface and the interior; (ii) the calculation, using Fourier transformation, of a correlation function that assesses the degree of molecular surface overlap and penetration upon relative shifts of the molecules in three dimensions; and (iii) a scan of the relative orientations of the molecules in three dimensions. The algorithm provides a list of correlation values indicating the extent of geometric match between the surfaces of the molecules; each of these values is associated with six numbers describing the relative position (translation and rotation) of the molecules. The procedure is thus equivalent to a six-dimensional search but much faster by design, and the computation time is only moderately dependent on molecular size. The procedure was tested and validated by using five known complexes for which the correct relative position of the molecules in the respective adducts was successfully predicted. The molecular pairs were deoxyhemoglobin and methemoglobin, tRNA synthetase-tyrosinyl adenylate, aspartic proteinase-peptide inhibitor, and trypsin-trypsin inhibitor. A more realistic test was performed with the last two pairs by using the structures of uncomplexed aspartic proteinase and trypsin inhibitor, respectively. The results are indicative of the extent of conformational changes in the molecules tolerated by the algorithm.

- Buttenschoen et al. (2024) — PoseBusters.

  PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Buttenschoen et al. (2024). *Chemical Science*. doi:10.1039/d3sc04185a

  The last few years have seen the development of numerous deep learning-based protein-ligand docking methods. They offer huge promise in terms of speed and accuracy. However, despite claims of state-of-the-art performance in terms of crystallographic root-mean-square deviation (RMSD), upon closer inspection, it has become apparent that they often produce physically implausible molecular structures. It is therefore not sufficient to evaluate these methods solely by RMSD to a native binding mode. It is vital, particularly for deep learning-based methods, that they are also evaluated on steric and energetic criteria. We present PoseBusters, a Python package that performs a series of standard quality checks using the well-established cheminformatics toolkit RDKit. The PoseBusters test suite validates chemical and geometric consistency of a ligand including its stereochemistry, and the physical plausibility of intra- and intermolecular measurements such as the planarity of aromatic rings, standard bond lengths, and protein-ligand clashes. Only methods that both pass these checks and predict native-like binding modes should be classed as having "state-of-the-art" performance. We use PoseBusters to compare five deep learning-based docking methods (DeepDock, DiffDock, EquiBind, TankBind, and Uni-Mol) and two well-established standard docking methods (AutoDock Vina and CCDC Gold) with and without an additional post-prediction energy minimisation step using a molecular mechanics force field. We show that both in terms of physical plausibility and the ability to generalise to examples that are distinct from the training data, no deep learning-based method yet outperforms classical docking tools. In addition, we find that molecular mechanics force fields contain docking-relevant physics missing from deep-learning methods. PoseBusters allows practitioners to assess docking and molecular generation methods and may inspire new inductive biases still required to improve deep learning-based methods, which will help drive the development of more accurate and more realistic predictions.

B.9.10 Protein design

- Kortemme (2024) — de novo protein design, from structures to functions.

  De novo protein design—From new structures to programmable functions. Kortemme (2024). *Cell*. doi:10.1016/j.cell.2023.12.028

  Methods from artificial intelligence (AI) trained on large datasets of sequences and structures can now "write" proteins with new shapes and molecular functions de novo, without starting from proteins found in nature. In this Perspective, I will discuss the state of the field of de novo protein design at the juncture of physics-based modeling approaches and AI. New protein folds and higher-order assemblies can be designed with considerable experimental success rates, and difficult problems requiring tunable control over protein conformations and precise shape complementarity for molecular recognition are coming into reach. Emerging approaches incorporate engineering principles-tunability, controllability, and modularity-into the design process from the beginning. Exciting frontiers lie in deconstructing cellular functions with de novo proteins and, conversely, constructing synthetic cellular signaling from the ground up. As methods improve, many more challenges are unsolved.

- Dahiyat and Mayo (1997) — fully automated design.

  De Novo Protein Design: Fully Automated Sequence Selection. Dahiyat & Mayo (1997). *Science*. doi:10.1126/science.278.5335.82

  > The first fully automated design and experimental validation of a novel sequence for an entire protein is described. A computational design algorithm based on physical chemical potential functions and stereochemical constraints was used to screen a combinatorial library of 1.9 x 10(27) possible amino acid sequences for compatibility with the design target, a betabetaalpha protein motif based on the polypeptide backbone structure of a zinc finger domain. A BLAST search shows that the designed sequence, full sequence design 1 (FSD-1), has very low identity to any known protein sequence. The solution structure of FSD-1 was solved by nuclear magnetic resonance spectroscopy and indicates that FSD-1 forms a compact well-ordered structure, which is in excellent agreement with the design target structure. This result demonstrates that computational methods can perform the immense combinatorial search required for protein design, and it suggests that an unbiased and quantitative algorithm can be used in various structural contexts.

- Kuhlman et al. (2003) — Top7, a new fold.

  Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Kuhlman et al. (2003). *Science*. doi:10.1126/science.1089427

  > A major challenge of computational protein design is the creation of novel proteins with arbitrarily chosen three-dimensional structures. Here, we used a general computational strategy that iterates between sequence design and structure prediction to design a 93-residue alpha/beta protein called Top7 with a novel sequence and topology. Top7 was found experimentally to be folded and extremely stable, and the x-ray crystal structure of Top7 is similar (root mean square deviation equals 1.2 angstroms) to the design model. The ability to design a new protein fold makes possible the exploration of the large regions of the protein universe not yet observed in nature.

- Koga et al. (2012), Jiang et al. (2008), ⋆ Huang et al. (2016) — ideal proteins, enzymes and the coming of age of design.

  Principles for designing ideal protein structures. Koga et al. (2012). *Nature*. doi:10.1038/nature11600

  > Unlike random heteropolymers, natural proteins fold into unique ordered structures. Understanding how these are encoded in amino-acid sequences is complicated by energetically unfavourable non-ideal features--for example kinked α-helices, bulged β-strands, strained loops and buried polar groups--that arise in proteins from evolutionary selection for biological function or from neutral drift. Here we describe an approach to designing ideal protein structures stabilized by completely consistent local and non-local interactions. The approach is based on a set of rules relating secondary structure patterns to protein tertiary motifs, which make possible the design of funnel-shaped protein folding energy landscapes leading into the target folded state. Guided by these rules, we designed sequences predicted to fold into ideal protein structures consisting of α-helices, β-strands and minimal loops. Designs for five different topologies were found to be monomeric and very stable and to adopt structures in solution nearly identical to the computational models. These results illuminate how the folding funnels of natural proteins arise and provide the foundation for engineering a new generation of functional proteins free from natural evolution.

  De Novo Computational Design of Retro-Aldol Enzymes. Jiang et al. (2008). *Science*. doi:10.1126/science.1152692

  > The creation of enzymes capable of catalyzing any desired chemical reaction is a grand challenge for computational protein design. Using new algorithms that rely on hashing techniques to construct active sites for multistep reactions, we designed retro-aldolases that use four different catalytic motifs to catalyze the breaking of a carbon-carbon bond in a nonnatural substrate. Of the 72 designs that were experimentally characterized, 32, spanning a range of protein folds, had detectable retro-aldolase activity. Designs that used an explicit water molecule to mediate proton shuffling were significantly more successful, with rate accelerations of up to four orders of magnitude and multiple turnovers, than those involving charged side-chain networks. The atomic accuracy of the design process was confirmed by the x-ray crystal structure of active designs embedded in two protein scaffolds, both of which were nearly superimposable on the design model.

The coming of age of de novo protein design. Huang et al. (2016). *Nature*. doi:10.1038/nature19946

> There are 20(200) possible amino-acid sequences for a 200-residue protein, of which the natural evolutionary process has sampled only an infinitesimal subset. De novo protein design explores the full sequence space, guided by the physical principles that underlie protein folding. Computational methodology has advanced to the point that a wide range of structures can be designed from scratch with atomic-level accuracy. Almost all protein engineering so far has involved the modification of naturally occurring proteins; it should now be possible to design new functional proteins from the ground up to tackle current challenges in biomedicine and nanotechnology.

- Dauparas et al. (2022) — ProteinMPNN.

Robust deep learning–based protein sequence design using ProteinMPNN. Dauparas et al. (2022). *Science*. doi:10.1126/science.add2187

> Although deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo protein designs have been generated using physically based approaches such as Rosetta. Here, we describe a deep learning-based protein sequence design method, ProteinMPNN, that has outstanding performance in both in silico and experimental tests. On native protein backbones, ProteinMPNN has a sequence recovery of 52.4% compared with 32.9% for Rosetta. The amino acid sequence at different positions can be coupled between single or multiple chains, enabling application to a wide range of current protein design challenges. We demonstrate the broad utility and high accuracy of ProteinMPNN using x-ray crystallography, cryo-electron microscopy, and functional studies by rescuing previously failed designs, which were made using Rosetta or AlphaFold, of protein monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target-binding proteins.

- Anishchenko et al. (2021), Wang et al. (2022) — hallucination.

De novo protein design by deep network hallucination. Anishchenko et al. (2021). *Nature*. doi:10.1038/s41586-021-04184-w

> There has been considerable recent progress in protein structure prediction using deep neural networks to predict inter-residue distances from amino acid sequences 1-3 . Here we investigate whether the information captured by such networks is sufficiently rich to generate new folded proteins with sequences unrelated to those of the naturally occurring proteins used in training the models. We generate random amino acid sequences, and input them into the trRosetta structure prediction network to predict starting residue-residue distance maps, which, as expected, are quite featureless. We then carry out Monte Carlo sampling in amino acid sequence space, optimizing the contrast (Kullback-Leibler divergence) between the inter-residue distance distributions predicted by the network and background distributions averaged over all proteins. Optimization from different random starting points resulted in novel proteins spanning a wide range of sequences and predicted structures. We obtained synthetic genes encoding 129 of the network-'hallucinated' sequences, and expressed and purified the proteins in Escherichia coli; 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. We determined the three-dimensional structures of three of the hallucinated proteins, two by X-ray crystallography and one by NMR, and these closely matched the hallucinated models. Thus, deep networks trained to predict native protein structures from their sequences can be inverted to design new proteins, and such networks and methods should contribute alongside traditional physics-based models to the de novo design of proteins with new functions.

Scaffolding protein functional sites using deep learning. Wang et al. (2022). *Science*. doi:10.1126/science.abn2100

> The binding and catalytic functions of proteins are generally mediated by a small number of functional residues held in place by the overall protein structure. Here, we describe deep learning approaches for scaffolding such functional sites without needing to prespecify the fold or secondary structure of the scaffold. The first approach, "constrained hallucination," optimizes sequences such that their predicted structures contain the desired functional site. The second approach, "inpainting," starts from the functional site and fills in additional sequence and structure to create a viable protein scaffold in a single forward pass through a specifically trained RoseTTAFold network. We use these two methods to design candidate immunogens, receptor traps, metalloproteins, enzymes, and protein-binding proteins and validate the designs using a combination of in silico and experimental tests.

- Pacesa et al. (2025) — BindCraft.

One-shot design of functional protein binders with BindCraft. Pacesa et al. (2025). *Nature*. doi:10.1038/s41586-025-09429-6

> Protein-protein interactions are at the core of all key biological processes. However, the complexity of the structural features that determine protein-protein interactions makes their design challenging. Here we present BindCraft, an open-source and automated pipeline for de novo protein binder design with experimental success rates of 10-100%. BindCraft leverages the weights of AlphaFold2 (ref. 1) to generate binders with nanomolar affinity without the need for high-throughput screening or experimental optimization, even in the absence of known binding sites. We successfully designed binders against a diverse set of challenging targets, including cell-surface receptors, common allergens, de novo designed proteins and multi-domain nucleases, such as CRISPR-Cas9. We showcase the functional and therapeutic potential of designed binders by reducing IgE binding to birch allergen in patient-derived samples, modulating Cas9 gene editing activity and reducing the cytotoxicity of a foodborne bacterial enterotoxin. Last, we use cell-surface-receptor-specific binders to redirect adeno-associated virus capsids for targeted gene delivery. This work represents a significant advancement towards a 'one design-one binder' approach in computational design, with immense potential in therapeutics, diagnostics and biotechnology.

- Ingraham et al. (2023), Krishna et al. (2024) — Chroma and RFdiffusion All-Atom.

Illuminating protein space with a programmable generative model. Ingraham et al. (2023). *Nature*. doi:10.1038/s41586-023-06728-8

> Three billion years of evolution has produced a tremendous diversity of protein molecules 1 , but the full potential of proteins is likely to be much greater. Accessing this potential has been challenging for both computation and experiments because the space of possible protein molecules is much larger than the space of those likely to have functions. Here we introduce Chroma, a generative model for proteins and protein complexes that can directly sample novel protein structures and sequences, and that can be conditioned to steer the generative process towards desired properties and functions. To enable this, we introduce a diffusion process that respects the conformational statistics of polymer ensembles, an efficient neural architecture for molecular systems that enables long-range reasoning with sub-quadratic scaling, layers for efficiently synthesizing three-dimensional structures of proteins from predicted inter-residue geometries and a general low-temperature sampling algorithm for diffusion models. Chroma achieves protein design as Bayesian inference under external constraints, which can involve symmetries, substructure, shape, semantics and even natural-language prompts. The experimental characterization of 310 proteins shows that sampling from Chroma results in proteins that are highly expressed, fold and have favourable biophysical properties. The crystal structures of two designed proteins exhibit atomistic agreement with Chroma samples (a backbone root-mean-square deviation of around 1.0 Å). With this unified approach to protein design, we hope to accelerate the programming of protein matter to benefit human health, materials science and synthetic biology.

Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Krishna et al. (2024). *Science*. doi:10.1126/science.adl2528

> Deep-learning methods have revolutionized protein structure prediction and design but are presently limited to protein-only systems. We describe RoseTTAFold All-Atom (RFAA), which combines a residue-based representation of amino acids and DNA bases with an atomic representation of all other groups to model assemblies that contain proteins, nucleic acids, small molecules, metals, and covalent modifications, given their sequences and chemical structures. By fine-tuning on denoising tasks, we developed RFdiffusion All-Atom (RFdiffusionAA), which builds protein structures around small molecules. Starting from random distributions of amino acid residues surrounding target small molecules, we designed and experimentally validated, through crystallography and binding measurements, proteins that bind the cardiac disease therapeutic digoxigenin, the enzymatic cofactor heme, and the light-harvesting molecule bilin.

- Cho et al. (2025) — Protein Hunter (preprint).

Protein Hunter: exploiting structure hallucination within diffusion for protein design. Cho et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.10.10.681530

1 Interactions between proteins and other biomolecules underlie nearly all biological processes, yet designing such interactions de novo remains challenging. Capturing their specific interactions and co-optimizing sequence and structure are difficult and often require extensive computation. We present Protein Hunter, a fast, fine-tuning-free framework for de novo protein design. Starting from an all-X sequence, we find diffusion-based structure prediction models hallucinate reasonable looking structures that can be further improved through iterative sequence re-design and structure re-prediction. This lightweight strategy achieves high AlphaFold3 in silico success rates across both unconditional and conditional generation tasks, including binders to proteins, cyclic peptides, small molecules, DNA, and RNA. Protein Hunter also supports multi-motif scaffolding and partial redesign, providing a general and efficient platform for de novo protein design across diverse molecular targets.

B.10 Protein function

Further papers for Chapter 10.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.10.1 Definitions and resources

- The Gene Ontology Consortium et al. (2023) — the Gene Ontology knowledgebase in 2023.

  The Gene Ontology knowledgebase in 2023. The Gene Ontology Consortium et al. (2023). *GENETICS*. doi:10.1093/genetics/iyad031

  > The Gene Ontology (GO) knowledgebase (http://geneontology.org) is a comprehensive resource concerning the functions of genes and gene products (proteins and noncoding RNAs). GO annotations cover genes from organisms across the tree of life as well as viruses, though most gene function knowledge currently derives from experiments carried out in a relatively small number of model organisms. Here, we provide an updated overview of the GO knowledgebase, as well as the efforts of the broad, international consortium of scientists that develops, maintains, and updates the GO knowledgebase. The GO knowledgebase consists of three components: (1) the GO-a computational knowledge structure describing the functional characteristics of genes; (2) GO annotations-evidence-supported statements asserting that a specific gene product has a particular functional characteristic; and (3) GO Causal Activity Models (GO-CAMs)-mechanistic models of molecular "pathways" (GO biological processes) created by linking multiple GO annotations using defined relations. Each of these components is continually expanded, revised, and updated in response to newly published discoveries and receives extensive QA checks, reviews, and user feedback. For each of these components, we provide a description of the current contents, recent developments to keep the knowledgebase up to date with new discoveries, and guidance on how users can best make use of the data that we provide. We conclude with future directions for the project.

- Chang et al. (2020), Bansal et al. (2021) — BRENDA and Rhea.

  BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Chang et al. (2020). *Nucleic Acids Research*. doi:10.1093/nar/gkaa1025

  > The BRENDA enzyme database (https://www.brenda-enzymes.org), established in 1987, has evolved into the main collection of functional enzyme and metabolism data. In 2018, BRENDA was selected as an ELIXIR Core Data Resource. BRENDA provides reliable data, continuous curation and updates of classified enzymes, and the integration of newly discovered enzymes. The main part contains >5 million data for ~90 000 enzymes from ~13 000 organisms, manually extracted from ~157 000 primary literature references, combined with information of text and data mining, data integration, and prediction algorithms. Supplements comprise disease-related data, protein sequences, 3D structures, genome annotations, ligand information, taxonomic, bibliographic, and kinetic data. BRENDA offers an easy access to enzyme information from quick to advanced searches, text- and structured-based queries for enzyme-ligand interactions, word maps, and visualization of enzyme data. The BRENDA Pathway Maps are completely revised and updated for an enhanced interactive and intuitive usability. The new design of the Enzyme Summary Page provides an improved access to each individual enzyme. A new protein structure 3D viewer was integrated. The prediction of the intracellular localization of eukaryotic enzymes has been

implemented. The new EnzymeDetector combines BRENDA enzyme annotations with protein and genome databases for the detection of eukaryotic and prokaryotic enzymes.

Rhea, the reaction knowledgebase in 2022. Bansal et al. (2021). *Nucleic Acids Research*. doi:10.1093/nar/gkab1016

> Rhea (https://www.rhea-db.org) is an expert-curated knowledgebase of biochemical reactions based on the chemical ontology ChEBI (Chemical Entities of Biological Interest) (https://www.ebi.ac.uk/chebi). In this paper, we describe a number of key developments in Rhea since our last report in the database issue of Nucleic Acids Research in 2019. These include improved reaction coverage in Rhea, the adoption of Rhea as the reference vocabulary for enzyme annotation in the UniProt knowledgebase UniProtKB (https://www.uniprot.org), the development of a new Rhea website, and the designation of Rhea as an ELIXIR Core Data Resource. We hope that these and other developments will enhance the utility of Rhea as a reference resource to study and engineer enzymes and the metabolic systems in which they function.

- Schnoes et al. (2013) — biases in experimental annotations.

Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. Schnoes et al. (2013). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1003063

> The ongoing functional annotation of proteins relies upon the work of curators to capture experimental findings from scientific literature and apply them to protein sequence and structure data. However, with the increasing use of high-throughput experimental assays, a small number of experimental studies dominate the functional protein annotations collected in databases. Here, we investigate just how prevalent is the "few articles - many proteins" phenomenon. We examine the experimentally validated annotation of proteins provided by several groups in the GO Consortium, and show that the distribution of proteins per published study is exponential, with 0.14% of articles providing the source of annotations for 25% of the proteins in the UniProt-GOA compilation. Since each of the dominant articles describes the use of an assay that can find only one function or a small group of functions, this leads to substantial biases in what we know about the function of many proteins. Mass-spectrometry, microscopy and RNAi experiments dominate high throughput experiments. Consequently, the functional information derived from these experiments is mostly of the subcellular location of proteins, and of the participation of proteins in embryonic developmental pathways. For some organisms, the information provided by different studies overlap by a large amount. We also show that the information provided by high throughput experiments is less specific than those provided by low throughput experiments. Given the experimental techniques available, certain biases in protein function annotation due to high-throughput experiments are unavoidable. Knowing that these biases exist and understanding their characteristics and extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

B.10.2 Homology and families

- Devos and Valencia (2000), Rost (2002), Tian and Skolnick (2003) — how well function is conserved between homologues.

Practical limits of function prediction. Devos & Valencia (2000). *Proteins: Structure, Function, and Genetics*. doi:10.1002/1097-0134(20001001)41:1<98::aid-prot120>3.0.co;2-s

> The widening gap between known protein sequences and their functions has led to the practice of assigning a potential function to a protein on the basis of sequence similarity to proteins whose function has been experimentally investigated. We present here a critical view of the theoretical and practical bases for this approach. The results obtained by analyzing a significant number of true sequence similarities, derived directly from structural alignments, point to the complexity of function prediction. Different aspects of protein function, including (i) enzymatic function classification, (ii) functional annotations in the form of key words, (iii) classes of cellular function, and (iv) conservation of binding sites can only be reliably transferred between similar sequences to a modest degree. The reason for this difficulty is a combination of the unavoidable database inaccuracies and the plasticity of protein function. In addition, analysis of the relationship between sequence and functional descriptions defines an empirical limit for pairwise-based functional annotations, namely, the three first digits of the six numbers used as descriptors of protein folds in the FSSP database can be predicted at an average level as low as 7.5% sequence identity,

two of the four EC digits at 15% identity, half of the SWISS-PROT key words related to protein function would require 20% identity, and the prediction of half of the residues in the binding site can be made at the 30% sequence identity level.

Enzyme Function Less Conserved than Anticipated. Rost (2002). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(02)00016-5

> The level of sequence similarity that implies similarity in protein structure is well established. Recently, many groups proposed thresholds for similarity in sequence implying similarity in enzymatic function. All previous results suggest the strong conservation of enzymatic function above levels of 50% pairwise sequence identity. Here, I argue that all groups substantially overestimated the conservation of enzyme function because their data sets were either too biased, or too small. An unbiased analysis suggested that less than 30% of the pair fragments above 50% sequence identity have entirely identical EC numbers. Another surprising finding was that even BLAST E-values below 10(-50) did not suffice to automatically transfer enzyme function without errors. As expected, most misclassifications originated from similarities in relatively short regions and/or from transferring annotations for different domains. Both problems cannot be corrected easily by adjusting the thresholds for automatic transfer of genome annotations. A score relating sequence identity to alignment length (distance from HSSP-threshold) outperformed statistical BLAST scores for high sequence similarity. In particular, the distance score allowed error-free transfer of enzyme function for the 10% most similar enzyme pairs. The results illustrated how difficult it is to assess the conservation of protein function and to guarantee error-free genome annotations, in general: sets with millions of pair comparisons might not suffice to arrive at statistically significant conclusions. In practice, the revised detailed estimates for the sequence conservation of enzyme function may provide important benchmarks for everyday sequence analysis and for more cautious automatic genome annotations.

How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?. Tian & Skolnick (2003). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2003.08.057

> Enzyme function conservation has been used to derive the threshold of sequence identity necessary to transfer function from a protein of known function to an unknown protein. Using pairwise sequence comparison, several studies suggested that when the sequence identity is above 40%, enzyme function is well conserved. In contrast, Rost argued that because of database bias, the results from such simple pairwise comparisons might be misleading. Thus, by grouping enzyme sequences into families based on sequence similarity and selecting representative sequences for comparison, he showed that enzyme function starts to diverge quickly when the sequence identity is below 70%. Here, we employ a strategy similar to Rost's to reduce the database bias; however, we classify enzyme families based not only on sequence similarity, but also on functional similarity, i.e. sequences in each family must have the same four digits or the same first three digits of the enzyme commission (EC) number. Furthermore, instead of selecting representative sequences for comparison, we calculate the function conservation of each enzyme family and then average the degree of enzyme function conservation across all enzyme families. Our analysis suggests that for functional transferability, 40% sequence identity can still be used as a confident threshold to transfer the first three digits of an EC number; however, to transfer all four digits of an EC number, above 60% sequence identity is needed to have at least 90% accuracy. Moreover, when PSI-BLAST is used, the magnitude of the E-value is found to be weakly correlated with the extent of enzyme function conservation in the third iteration of PSI-BLAST. As a result, functional annotation based on the E-values from PSI-BLAST should be used with caution. We also show that by employing an enzyme family-specific sequence identity threshold above which 100% functional conservation is required, functional inference of unknown sequences can be accurately accomplished. However, this comes at a cost: those true positive sequences below this threshold cannot be uniquely identified.

- Gilks et al. (2002), Schnoes et al. (2009) — error propagation and misannotation.

Modeling the percolation of annotation errors in a database of protein sequences. Gilks et al. (2002). *Bioinformatics*. doi:10.1093/bioinformatics/18.12.1641

> Public sequence databases contain information on the sequence, structure and function of proteins. Genome sequencing projects have led to a rapid increase in protein sequence information, but reliable, experimentally verified, information on protein function lags a long way behind. To address this deficit, functional annotation in protein databases is often inferred by sequence similarity to homologous, annotated proteins, with the attendant possibility of error. Now, the functional annotation in these homologous proteins may itself have been acquired through sequence similarity to yet other proteins, and it is generally not possible to determine how the functional annotation of any given protein has been

acquired. Thus the possibility of chains of misannotation arises, a process we term 'error percolation'. With some simple assumptions, we develop a dynamical probabilistic model for these misannotation chains. By exploring the consequences of the model for annotation quality it is evident that this iterative approach leads to a systematic deterioration of database quality.

Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. Schnoes et al. (2009). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000605

> Due to the rapid release of new data from genome sequencing projects, the majority of protein sequences in public databases have not been experimentally characterized; rather, sequences are annotated using computational analysis. The level of misannotation and the types of misannotation in large public databases are currently unknown and have not been analyzed in depth. We have investigated the misannotation levels for molecular function in four public protein sequence databases (UniProtKB/Swiss-Prot, GenBank NR, UniProtKB/TrEMBL, and KEGG) for a model set of 37 enzyme families for which extensive experimental information is available. The manually curated database Swiss-Prot shows the lowest annotation error levels (close to 0% for most families); the two other protein sequence databases (GenBank NR and TrEMBL) and the protein sequences in the KEGG pathways database exhibit similar and surprisingly high levels of misannotation that average 5%-63% across the six superfamilies studied. For 10 of the 37 families examined, the level of misannotation in one or more of these databases is >80%. Examination of the NR database over time shows that misannotation has increased from 1993 to 2005. The types of misannotation that were found fall into several categories, most associated with "overprediction" of molecular function. These results suggest that misannotation in enzyme superfamilies containing multiple families that catalyze different reactions is a larger problem than has been recognized. Strategies are suggested for addressing some of the systematic problems contributing to these high levels of misannotation.

- Nehrt et al. (2011), Altenhoff et al. (2012) — the ortholog conjecture.

Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. Nehrt et al. (2011). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002073

> A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the "ortholog conjecture"). Many methods used to computationally predict protein function are based on this assumption, even though it is largely untested. Here we present the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We use the experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities. Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. We also find that paralogous pairs residing on the same chromosome are more functionally similar than those on different chromosomes, perhaps due to higher levels of interlocus gene conversion between these pairs. In addition to offering implications for the computational prediction of protein function, our results shed light on the relationship between sequence divergence and functional divergence. We conclude that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act.

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. Altenhoff et al. (2012). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002514

> The function of most proteins is not determined experimentally, but is extrapolated from homologs. According to the "ortholog conjecture", or standard model of phylogenomics, protein function changes rapidly after duplication, leading to paralogs with different functions, while orthologs retain the ancestral function. We report here that a comparison of experimentally supported functional annotations among homologs from 13 genomes mostly supports this model. We show that to analyze GO annotation effectively, several confounding factors need to be controlled: authorship bias, variation of GO term frequency among species, variation of background similarity among species pairs, and propagated annotation bias. After controlling for these biases, we observe that orthologs have generally more similar functional annotations than paralogs. This is especially strong for sub-cellular localization. We observe only a weak decrease in functional similarity with increasing sequence divergence. These findings hold over a large diversity of

species; notably orthologs from model organisms such as E. coli, yeast or mouse have conserved function with human proteins.

- Cantalapiedra et al. (2021) — eggNOG-mapper.

  eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Cantalapiedra et al. (2021). *Molecular Biology and Evolution*. doi:10.1093/molbev/msab293

  > Even though automated functional annotation of genes represents a fundamental step in most genomic and metagenomic workflows, it remains challenging at large scales. Here, we describe a major upgrade to eggNOG-mapper, a tool for functional annotation based on precomputed orthology assignments, now optimized for vast (meta)genomic data sets. Improvements in version 2 include a full update of both the genomes and functional databases to those from eggNOG v5, as well as several efficiency enhancements and new features. Most notably, eggNOG-mapper v2 now allows for: 1) de novo gene prediction from raw contigs, 2) built-in pairwise orthology prediction, 3) fast protein domain discovery, and 4) automated GFF decoration. eggNOG-mapper v2 is available as a standalone tool or as an online service at http://eggnog-mapper.embl.de.

- Blum et al. (2024), Bileschi et al. (2022) — InterPro and deep-learning family classification.

  InterPro: the protein sequence classification resource in 2025. Blum et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1082

  > InterPro (https://www.ebi.ac.uk/interpro) is a freely accessible resource for the classification of protein sequences into families. It integrates predictive models, known as signatures, from multiple member databases to classify sequences into families and predict the presence of domains and significant sites. The InterPro database provides annotations for over 200 million sequences, ensuring extensive coverage of UniProtKB, the standard repository of protein sequences, and includes mappings to several other major resources, such as Gene Ontology (GO), Protein Data Bank in Europe (PDBe) and the AlphaFold Protein Structure Database. In this publication, we report on the status of InterPro (version 101.0), detailing new developments in the database, associated web interface and software. Notable updates include the increased integration of structures predicted by AlphaFold and the enhanced description of protein families using artificial intelligence. Over the past two years, more than 5000 new InterPro entries have been created. The InterPro website now offers access to 85 000 protein families and domains from its member databases and serves as a long-term archive for retired databases. InterPro data, software and tools are freely available.

  Using deep learning to annotate the protein universe. Bileschi et al. (2022). *Nature Biotechnology*. doi:10.1038/s41587-021-01179-w

  > Understanding the relationship between amino acid sequence and protein function is a long-standing challenge with far-reaching scientific and translational implications. State-of-the-art alignment-based techniques cannot predict function for one-third of microbial protein sequences, hampering our ability to exploit data from diverse organisms. Here, we train deep learning models to accurately predict functional annotations for unaligned amino acid sequences across rigorous benchmark assessments built from the 17,929 families of the protein families database Pfam. The models infer known patterns of evolutionary substitutions and learn representations that accurately cluster sequences from unseen families. Combining deep models with existing methods significantly improves remote homology detection, suggesting that the deep models learn complementary information. This approach extends the coverage of Pfam by >9.5%, exceeding additions made over the last decade, and predicts function for 360 human reference proteome proteins with no previous Pfam annotation. These results suggest that deep learning models will be a core component of future protein annotation tools.

B.10.3 Genomic context and networks

- Marcotte, Pellegrini, Thompson, et al. (1999) — a combined algorithm for genome-wide function prediction.

  A combined algorithm for genome-wide prediction of protein function. Marcotte et al. (1999). *Nature*. doi:10.1038/47048

> The availability of over 20 fully sequenced genomes has driven the development of new methods to find protein function and interactions. Here we group proteins by correlated evolution, correlated messenger RNA expression patterns and patterns of domain fusion to determine functional relationships among the 6,217 proteins of the yeast Saccharomyces cerevisiae. Using these methods, we discover over 93,000 pairwise links between functionally related yeast proteins. Links between characterized and uncharacterized proteins allow a general function to be assigned to more than half of the 2,557 previously uncharacterized yeast proteins. Examples of functional links are given for a protein family of previously unknown function, a protein whose human homologues are implicated in colon cancer and the yeast prion Sup35.

- Enright and Ouzounis (2001), Tsoka and Ouzounis (2000), Snel et al. (2000), R. F. Doolittle (1999) — gene fusion and fission.

  Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Enright & Ouzounis (2001). *Genome Biology*. doi:10.1186/gb-2001-2-9-research0034

  > Background It has recently been shown that the detection of gene fusion events across genomes can be used for predicting functional associations of proteins, including physical interaction or complex formation. To obtain such predictions we have made an exhaustive search for gene fusion events within 24 available completely sequenced genomes. Results Each genome was used as a query against the remaining 23 complete genomes to detect gene fusion events. Using an improved, fully automatic protocol, a total of 7,224 single-domain proteins that are components of gene fusions in other genomes were detected, many of which were identified for the first time. The total number of predicted pairwise functional associations is 39,730 for all genomes. Component pairs were identified by virtue of their similarity to 2,365 multidomain composite proteins. We also show for the first time that gene fusion is a complex evolutionary process with a number of contributory factors, including paralogy, genome size and phylogenetic distance. On average, 9% of genes in a given genome appear to code for single-domain, component proteins predicted to be functionally associated. These proteins are detected by an additional 4% of genes that code for fused, composite proteins. Conclusions These results provide an exhaustive set of functionally associated genes and also delineate the power of fusion analysis for the prediction of protein interactions.

  Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Tsoka & Ouzounis (2000). *Nature Genetics*. doi:10.1038/79847

  Genome evolution. Snel et al. (2000). *Trends in Genetics*. doi:10.1016/s0168-9525(99)01924-1

  Do you dig my groove?. Doolittle (1999). *Nature Genetics*. doi:10.1038/12597

- Dandekar (1998), Overbeek et al. (1999) — conserved gene order.

  Conservation of gene order: a fingerprint of proteins that physically interact. Dandekar (1998). *Trends in Biochemical Sciences*. doi:10.1016/s0968-0004(98)01274-2

  > A systematic comparison of nine bacterial and archaeal genomes reveals a low level of gene-order (and operon architecture) conservation. Nevertheless, a number of gene pairs are conserved. The proteins encoded by conserved gene pairs appear to interact physically. This observation can therefore be used to predict functions of, and interactions between, prokaryotic gene products.

  The use of gene clusters to infer functional coupling. Overbeek et al. (1999). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.96.6.2896

  > Previously, we presented evidence that it is possible to predict functional coupling between genes based on conservation of gene clusters between genomes. With the rapid increase in the availability of prokaryotic sequence data, it has become possible to verify and apply the technique. In this paper, we extend our characterization of the parameters that determine the utility of the approach, and we generalize the approach in a way that supports detection of common classes of functionally coupled genes (e.g., transport and signal transduction clusters). Now that the analysis includes over 30 complete or nearly complete genomes, it has become clear that this approach will play a significant role in supporting efforts to assign functionality to the remaining uncharacterized genes in sequenced genomes.

- Pellegrini et al. (1999) — phylogenetic profiles.

  Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Pellegrini et al. (1999). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.96.8.4285

  > Determining protein functions from genomic sequences is a central goal of bioinformatics. We present a method based on the assumption that proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion. During evolution, all such functionally linked proteins tend to be either preserved or eliminated in a new species. We describe this property of correlated evolution by characterizing each protein by its phylogenetic profile, a string that encodes the presence or absence of a protein in every known genome. We show that proteins having matching or similar profiles strongly tend to be functionally linked. This method of phylogenetic profiling allows us to predict the function of uncharacterized proteins.

- Sharan et al. (2007) — network-based prediction of protein function.

  Network-based prediction of protein function. Sharan et al. (2007). *Molecular Systems Biology*. doi:10.1038/msb4100129

  > Functional annotation of proteins is a fundamental problem in the post-genomic era. The recent availability of protein interaction networks for many model species has spurred on the development of computational methods for interpreting such data in order to elucidate protein function. In this review, we describe the current computational approaches for the task, including direct methods, which propagate functional information through the network, and module-assisted methods, which infer functional modules within the network and use those for the annotation task. Although a broad variety of interesting approaches has been developed, further progress in the field will depend on systematic evaluation of the methods and their dissemination in the biological community.

- ⋆ Garzón et al. (2016) — a computational interactome and functional annotation of the human proteome.

  A computational interactome and functional annotation for the human proteome. Garzón et al. (2016). *eLife*. doi:10.7554/elife.18715

  > We present a database, PrePPI (Predicting Protein-Protein Interactions), of more than 1.35 million predicted protein-protein interactions (PPIs). Of these at least 127,000 are expected to constitute direct physical interactions although the actual number may be much larger (~500,000). The current PrePPI, which contains predicted interactions for about 85% of the human proteome, is related to an earlier version but is based on additional sources of interaction evidence and is far larger in scope. The use of structural relationships allows PrePPI to infer numerous previously unreported interactions. PrePPI has been subjected to a series of validation tests including reproducing known interactions, recapitulating multi-protein complexes, analysis of disease associated SNPs, and identifying functional relationships between interacting proteins. We show, using Gene Set Enrichment Analysis (GSEA), that predicted interaction partners can be used to annotate a protein's function. We provide annotations for most human proteins, including many annotated as having unknown function.

B.10.4 Structure and machine learning

- Laskowski et al. (2005) — ProFunc: function from 3D structure.

  ProFunc: a server for predicting protein function from 3D structure. Laskowski et al. (2005). *Nucleic Acids Research*. doi:10.1093/nar/gki414

  > ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) is a web server for predicting the likely function of proteins whose 3D structure is known but whose function is not. Users submit the coordinates of their structure to the server in PDB format. ProFunc makes use of both existing and novel methods to analyse the protein's sequence and structure identifying functional motifs or close relationships to functionally characterized proteins. A summary of the analyses provides an at-a-glance view of what each of the different methods has found. More detailed results are available on separate pages. Often where one method has failed to find anything useful another may be more forthcoming. The server is likely

to be of most use in structural genomics where a large proportion of the proteins whose structures are solved are of hypothetical proteins of unknown function. However, it may also find use in a comparative analysis of members of large protein families. It provides a convenient compendium of sequence and structural information that often hold vital functional clues to be followed up experimentally.

- Gligorijević et al. (2021) — DeepFRI: graph convolutional networks on structures.

  Structure-based protein function prediction using graph convolutional networks. Gligorijević et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-23303-9

  > The rapid increase in the number of proteins in sequence databases and the diversity of their functions challenge computational approaches for automated function prediction. Here, we introduce DeepFRI, a Graph Convolutional Network for predicting protein functions by leveraging sequence features extracted from a protein language model and protein structures. It outperforms current leading methods and sequence-based Convolutional Neural Networks and scales to the size of current sequence repositories. Augmenting the training set of experimental structures with homology models allows us to significantly expand the number of predictable functions. DeepFRI has significant de-noising capability, with only a minor drop in performance when experimental structures are replaced by protein models. Class activation mapping allows function predictions at an unprecedented resolution, allowing site-specific annotations at the residue-level in an automated manner. We show the utility and high performance of our method by annotating structures from the PDB and SWISS-MODEL, making several new confident function predictions. DeepFRI is available as a webserver at https://beta.deepfri.flatironinstitute.org/ .

- Kulmanov and Hoehndorf (2019) — DeepGOPlus.

  DeepGOPlus: improved protein function prediction from sequence. Kulmanov & Hoehndorf (2019). *Bioinformatics*. doi:10.1093/bioinformatics/btz595

  > Motivation Protein function prediction is one of the major tasks of bioinformatics that can help in wide range of biological problems such as understanding disease mechanisms or finding drug targets. Many methods are available for predicting protein functions from sequence based features, protein-protein interaction networks, protein structure or literature. However, other than sequence, most of the features are difficult to obtain or not available for many proteins thereby limiting their scope. Furthermore, the performance of sequence-based function prediction methods is often lower than methods that incorporate multiple features and predicting protein functions may require a lot of time. Results We developed a novel method for predicting protein functions from sequence alone which combines deep convolutional neural network (CNN) model with sequence similarity based predictions. Our CNN model scans the sequence for motifs which are predictive for protein functions and combines this with functions of similar proteins (if available). We evaluate the performance of DeepGOPlus using the CAFA3 evaluation measures and achieve an Fmax of 0.390, 0.557 and 0.614 for BPO, MFO and CCO evaluations, respectively. These results would have made DeepGOPlus one of the three best predictors in CCO and the second best performing method in the BPO and MFO evaluations. We also compare DeepGOPlus with state-of-the-art methods such as DeepText2GO and GOLabeler on another dataset. DeepGOPlus can annotate around 40 protein sequences per second on common hardware, thereby making fast and accurate function predictions available for a wide range of proteins. Availability and implementation http://deepgoplus.bio2vec.net/ . Supplementary information Supplementary data are available at Bioinformatics online.

- Littmann et al. (2021) — transferring GO annotations with embeddings.

  Embeddings from deep learning transfer GO annotations beyond homology. Littmann et al. (2021). *Scientific Reports*. doi:10.1038/s41598-020-80786-0

  > Knowing protein function is crucial to advance molecular and medical biology, yet experimental function annotations through the Gene Ontology (GO) exist for fewer than 0.5% of all known proteins. Computational methods bridge this sequence-annotation gap typically through homology-based annotation transfer by identifying sequence-similar proteins with known function or through prediction methods using evolutionary information. Here, we propose predicting GO terms through annotation transfer based on proximity of proteins in the SeqVec embedding rather than in sequence space. These embeddings originate from deep learned language models (LMs) for protein sequences (SeqVec) transferring the knowledge gained from predicting the next amino acid in 33 million protein sequences. Replicating the conditions of CAFA3, our method reaches an F max of 37 ± 2%, 50 ± 3%, and 57 ± 2% for BPO, MFO, and CCO, respectively. Numerically, this appears close to the top ten CAFA3 methods. When restricting the annotation transfer to proteins with max BPO 33 ± 2%, MFO 43 ± 3%, CCO 53 ± 2%); this still outperforms

naïve sequence-based transfer. Preliminary results from CAFA4 appear to confirm these findings. Overall, this new concept is likely to change the annotation of proteins, in particular for proteins from smaller families or proteins with intrinsically disordered regions.

- Yu et al. (2023) — CLEAN: enzyme function prediction with contrastive learning.

  Enzyme function prediction using contrastive learning. Yu et al. (2023). *Science*. doi:10.1126/science.adf2465

  > Enzyme function annotation is a fundamental challenge, and numerous computational tools have been developed. However, most of these tools cannot accurately predict functional annotations, such as enzyme commission (EC) number, for less-studied proteins or those with previously uncharacterized functions or multiple activities. We present a machine learning algorithm named CLEAN (contrastive learning-enabled enzyme annotation) to assign EC numbers to enzymes with better accuracy, reliability, and sensitivity compared with the state-of-the-art tool BLASTp. The contrastive learning framework empowers CLEAN to confidently (i) annotate understudied enzymes, (ii) correct mislabeled enzymes, and (iii) identify promiscuous enzymes with two or more EC numbers-functions that we demonstrate by systematic in silico and in vitro experiments. We anticipate that this tool will be widely used for predicting the functions of uncharacterized enzymes, thereby advancing many fields, such as genomics, synthetic biology, and biocatalysis.

B.10.5 Localisation

- Blobel and Dobberstein (1975) — the signal hypothesis.

  Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma.. Blobel & Dobberstein (1975). *The Journal of cell biology*. doi:10.1083/jcb.67.3.835

  > Fractionation of MOPC 41 DL-1 tumors revealed that the mRNA for the light chain of immunoglobulin is localized exclusively in membrane-bound ribosomes. It was shown that the translation product of isolated light chain mRNA in a heterologous protein-synthesizing system in vitro is larger than the authentic secreted light chain; this confirms similar results from several laboratories. The synthesis in vitro of a precursor protein of the light chain is not an artifact of translation in a heterologous system, because it was shown that detached polysomes, isolated from detergent-treated rough microsomes, not only contain nascent light chains which have already been proteolytically processed in vivo but also contain unprocessed nascent light chains. In vitro completion of these nascent light chains thus resulted in the synthesis of some chains having the same mol wt as the authentic secreted light chains, because of completion of in vivo proteolytically processed chains and of other chains which, due to the completion of unprocessed chains, have the same mol wt as the precursor of the light chain. In contrast, completion of the nascent light chains contained in rough microsomes resulted in the synthesis of only processed light chains. Taken together, these results indicate that the processing activity is present in isolated rough microsomes, that it is localized in the membrane moiety of rough microsomes, and, therefore, that it was most likely solubilized during detergent treatment used for the isolation of detached polysomes. Furthermore, these results established that processing in vivo takes place before completion of the nascent chain. The data also indicate that in vitro processing of nascent chains by rough microsomes is dependent on ribosome binding to the membrane. If the latter process is interfered with by aurintricarboxylic acid, rough microsomes also synthesize some unprocessed chains. The data presented in this paper have been interpreted in the light of a recently proposed hypothesis. This hypothesis, referred to as the signal hypothesis, is described in greater detail in the Discussion section.

- Heijne (1986) — predicting signal-sequence cleavage sites.

  A new method for predicting signal sequence cleavage sites. von Heijne (1986). *Nucleic Acids Research*. doi:10.1093/nar/14.11.4683

  > A new method for identifying secretory signal sequences and for predicting the site of cleavage between a signal sequence and the mature exported protein is described. The predictive accuracy is estimated to be around 75-80% for both prokaryotic and eukaryotic proteins.

- Nielsen et al. (1997), Petersen et al. (2011) — SignalP and the transmembrane problem.

  Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Nielsen et al. (1997). *Protein Engineering Design and Selection*. doi:10.1093/protein/10.1.1

  > We have developed a new method for the identification of signal peptides and their cleavage sites based on neural networks trained on separate sets of prokaryotic and eukaryotic sequence. The method performs significantly better than previous prediction schemes and can easily be applied on genome-wide data sets. Discrimination between cleaved signal peptides and uncleaved N-terminal signal-anchor sequences is also possible, though with lower precision. Predictions can be made on a publicly available WWW server.

  SignalP 4.0: discriminating signal peptides from transmembrane regions. Petersen et al. (2011). *Nature Methods*. doi:10.1038/nmeth.1701

- Nakai and Horton (1999) — PSORT.

  PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Nakai & Horton (1999). *Trends in Biochemical Sciences*. doi:10.1016/s0968-0004(98)01336-x

- Emanuelsson et al. (2000), Almagro Armenteros et al. (2019) — TargetP.

  Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. Emanuelsson et al. (2000). *Journal of Molecular Biology*. doi:10.1006/jmbi.2000.3903

  > A neural network-based tool, TargetP, for large-scale subcellular location prediction of newly identified proteins has been developed. Using N-terminal sequence information only, it discriminates between proteins destined for the mitochondrion, the chloroplast, the secretory pathway, and "other" localizations with a success rate of 85% (plant) or 90% (non-plant) on redundancy-reduced test sets. From a TargetP analysis of the recently sequenced Arabidopsis thaliana chromosomes 2 and 4 and the Ensembl Homo sapiens protein set, we estimate that 10% of all plant proteins are mitochondrial and 14% chloroplastic, and that the abundance of secretory proteins, in both Arabidopsis and Homo, is around 10%. TargetP also predicts cleavage sites with levels of correctly predicted sites ranging from approximately 40% to 50% (chloroplastic and mitochondrial presequences) to above 70% (secretory signal peptides). TargetP is available as a web-server at http://www.cbs.dtu.dk/services/TargetP/.

  Detecting sequence signals in targeting peptides using deep learning. Almagro Armenteros et al. (2019). *Life Science Alliance*. doi:10.26508/lsa.201900429

  > In bioinformatics, machine learning methods have been used to predict features embedded in the sequences. In contrast to what is generally assumed, machine learning approaches can also provide new insights into the underlying biology. Here, we demonstrate this by presenting TargetP 2.0, a novel state-of-the-art method to identify N-terminal sorting signals, which direct proteins to the secretory pathway, mitochondria, and chloroplasts or other plastids. By examining the strongest signals from the attention layer in the network, we find that the second residue in the protein, that is, the one following the initial methionine, has a strong influence on the classification. We observe that two-thirds of chloroplast and thylakoid transit peptides have an alanine in position 2, compared with 20% in other plant proteins. We also note that in fungi and single-celled eukaryotes, less than 30% of the targeting peptides have an amino acid that allows the removal of the N-terminal methionine compared with 60% for the proteins without targeting peptide. The importance of this feature for predictions has not been highlighted before.

- ⋆ Almagro Armenteros et al. (2017), Thumuluri et al. (2022) — DeepLoc and DeepLoc 2.0.

  DeepLoc: prediction of protein subcellular localization using deep learning. Almagro Armenteros et al. (2017). *Bioinformatics*. doi:10.1093/bioinformatics/btx431

  > Motivation The prediction of eukaryotic protein subcellular localization is a well-studied topic in bioinformatics due to its relevance in proteomics research. Many machine learning methods have been successfully applied in this task, but in most of them, predictions rely on annotation of homologues from knowledge databases. For novel proteins where no annotated homologues exist, and for predicting the effects of sequence variants, it is desirable to have methods for predicting protein properties from sequence information only. Results Here, we present a prediction algorithm using deep neural networks to predict

protein subcellular localization relying only on sequence information. At its core, the prediction model uses a recurrent neural network that processes the entire protein sequence and an attention mechanism identifying protein regions important for the subcellular localization. The model was trained and tested on a protein dataset extracted from one of the latest UniProt releases, in which experimentally annotated proteins follow more stringent criteria than previously. We demonstrate that our model achieves a good accuracy (78% for 10 categories; 92% for membrane-bound or soluble), outperforming current state-of-the-art algorithms, including those relying on homology information. Availability and implementation The method is available as a web server at http://www.cbs.dtu.dk/services/DeepLoc. Example code is available at https://github.com/JJAlmagro/subcellular_localization. The dataset is available at http://www.cbs.dtu.dk/services/DeepLoc/data.php. Contact jjalma@dtu.dk.

DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Thumuluri et al. (2022). *Nucleic Acids Research*. doi:10.1093/nar/gkac278

The prediction of protein subcellular localization is of great relevance for proteomics research. Here, we propose an update to the popular tool DeepLoc with multi-localization prediction and improvements in both performance and interpretability. For training and validation, we curate eukaryotic and human multi-location protein datasets with stringent homology partitioning and enriched with sorting signal information compiled from the literature. We achieve state-of-the-art performance in DeepLoc 2.0 by using a pre-trained protein language model. It has the further advantage that it uses sequence input rather than relying on slower protein profiles. We provide two means of better interpretability: an attention output along the sequence and highly accurate prediction of nine different types of protein sorting signals. We find that the attention output correlates well with the position of sorting signals. The webserver is available at services.healthtech.dtu.dk/service.php?DeepLoc-2.0.

B.10.6 Variant effects

- Ng (2003), Adzhubei et al. (2010), Kircher et al. (2014) — SIFT, PolyPhen-2 and CADD.

SIFT: predicting amino acid changes that affect protein function. Ng (2003). *Nucleic Acids Research*. doi:10.1093/nar/gkg509

Single nucleotide polymorphism (SNP) studies and random mutagenesis projects identify amino acid substitutions in protein-coding regions. Each substitution has the potential to affect protein function. SIFT (Sorting Intolerant From Tolerant) is a program that predicts whether an amino acid substitution affects protein function so that users can prioritize substitutions for further study. We have shown that SIFT can distinguish between functionally neutral and deleterious amino acid changes in mutagenesis studies and on human polymorphisms. SIFT is available at http://blocks.fhcrc.org/sift/SIFT.html.

A method and server for predicting damaging missense mutations. Adzhubei et al. (2010). *Nature Methods*. doi:10.1038/nmeth0410-248

A general framework for estimating the relative pathogenicity of human genetic variants. Kircher et al. (2014). *Nature Genetics*. doi:10.1038/ng.2892

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for example, to missense changes). Here we describe Combined Annotation-Dependent Depletion (CADD), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant. We implement CADD as a support vector machine trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. We precompute C scores for all 8.6 billion possible human single-nucleotide variants and enable scoring of short insertions-deletions. C scores correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, and they highly rank known pathogenic variants within individual genomes. The ability of CADD to prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures is unmatched by any current single-annotation method.

- Grimm et al. (2015) — circularity in the evaluation of variant predictors.

  The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. Grimm et al. (2015). *Human Mutation*. doi:10.1002/humu.22768

  > Prioritizing missense variants for further experimental investigation is a key challenge in current sequencing studies for exploring complex and Mendelian diseases. A large number of in silico tools have been employed for the task of pathogenicity prediction, including PolyPhen-2, SIFT, FatHMM, MutationTaster-2, MutationAssessor, Combined Annotation Dependent Depletion, LRT, phyloP, and GERP++, as well as optimized methods of combining tool scores, such as Condel and Logit. Due to the wealth of these methods, an important practical question to answer is which of these tools generalize best, that is, correctly predict the pathogenic character of new variants. We here demonstrate in a study of 10 tools on five datasets that such a comparative evaluation of these tools is hindered by two types of circularity: they arise due to (1) the same variants or (2) different variants from the same protein occurring both in the datasets used for training and for evaluation of these tools, which may lead to overly optimistic results. We show that comparative evaluations of predictors that do not address these types of circularity may erroneously conclude that circularity confounded tools are most accurate among all tools, and may even outperform optimized combinations of tools.

- Hopf et al. (2017), Riesselman et al. (2018), Frazer et al. (2021) — EVmutation, DeepSequence and EVE.

  Mutation effects predicted from sequence co-variation. Hopf et al. (2017). *Nature Biotechnology*. doi:10.1038/nbt.3769

  > Many high-throughput experimental technologies have been developed to assess the effects of large numbers of mutations (variation) on phenotypes. However, designing functional assays for these methods is challenging, and systematic testing of all combinations is impossible, so robust methods to predict the effects of genetic variation are needed. Most prediction methods exploit evolutionary sequence conservation but do not consider the interdependencies of residues or bases. We present EVmutation, an unsupervised statistical method for predicting the effects of mutations that explicitly captures residue dependencies between positions. We validate EVmutation by comparing its predictions with outcomes of high-throughput mutagenesis experiments and measurements of human disease mutations and show that it outperforms methods that do not account for epistasis. EVmutation can be used to assess the quantitative effects of mutations in genes of any organism. We provide pre-computed predictions for ~7,000 human proteins at http://evmutation.org/.

  Deep generative models of genetic variation capture the effects of mutations. Riesselman et al. (2018). *Nature Methods*. doi:10.1038/s41592-018-0138-4

  > The functions of proteins and RNAs are defined by the collective interactions of many residues, and yet most statistical models of biological sequences consider sites nearly independently. Recent approaches have demonstrated benefits of including interactions to capture pairwise covariation, but leave higher-order dependencies out of reach. Here we show how it is possible to capture higher-order, context-dependent constraints in biological sequences via latent variable models with nonlinear dependencies. We found that DeepSequence (https://github.com/debbiemarkslab/DeepSequence), a probabilistic model for sequence families, predicted the effects of mutations across a variety of deep mutational scanning experiments substantially better than existing methods based on the same evolutionary data. The model, learned in an unsupervised manner solely on the basis of sequence information, is grounded with biologically motivated priors, reveals the latent organization of sequence families, and can be used to explore new parts of sequence space.

  Disease variant prediction with deep generative models of evolutionary data. Frazer et al. (2021). *Nature*. doi:10.1038/s41586-021-04043-8

  > Quantifying the pathogenicity of protein variants in human disease-related genes would have a marked effect on clinical decisions, yet the overwhelming majority (over 98%) of these variants still have unknown consequences 1-3 . In principle, computational methods could support the large-scale interpretation of genetic variants. However, state-of-the-art methods 4-10 have relied on training machine learning models on known disease labels. As these labels are sparse, biased and of variable quality, the resulting models have been considered insufficiently reliable 11 . Here we propose an approach that leverages deep generative models to predict variant pathogenicity without relying on labels. By modelling the distribution of sequence variation across organisms, we implicitly capture constraints on the protein sequences

> that maintain fitness. Our model EVE (evolutionary model of variant effect) not only outperforms computational approaches that rely on labelled data but also performs on par with, if not better than, predictions from high-throughput experiments, which are increasingly used as evidence for variant classification 12-16 . We predict the pathogenicity of more than 36 million variants across 3,219 disease genes and provide evidence for the classification of more than 256,000 variants of unknown significance. Our work suggests that models of evolutionary information can provide valuable independent evidence for variant interpretation that will be widely useful in research and clinical settings.

- Meier et al. (2021), Brandes et al. (2023) — variant effects from protein language models.

  Language models enable zero-shot prediction of the effects of mutations on protein function. Meier et al. (2021). *bioRxiv (preprint)*. doi:10.1101/2021.07.09.450648

  > Modeling the effect of sequence variation on function is a fundamental problem for understanding and designing proteins. Since evolution encodes information about function into patterns in protein sequences, unsupervised models of variant effects can be learned from sequence data. The approach to date has been to fit a model to a family of related sequences. The conventional setting is limited, since a new model must be trained for each prediction task. We show that using only zero-shot inference, without any supervision from experimental data or additional training, protein language models capture the functional effects of sequence variation, performing at state-of-the-art.

  Genome-wide prediction of disease variant effects with a deep protein language model. Brandes et al. (2023). *Nature Genetics*. doi:10.1038/s41588-023-01465-0

  > Predicting the effects of coding variants is a major challenge. While recent deep-learning models have improved variant effect prediction accuracy, they cannot analyze all coding variants due to dependency on close homologs or software limitations. Here we developed a workflow using ESM1b, a 650-million-parameter protein language model, to predict all ~450 million possible missense variant effects in the human genome, and made all predictions available on a web portal. ESM1b outperformed existing methods in classifying ~150,000 ClinVar/HGMD missense variants as pathogenic or benign and predicting measurements across 28 deep mutational scan datasets. We further annotated ~2 million variants as damaging only in specific protein isoforms, demonstrating the importance of considering all isoforms when predicting variant effects. Our approach also generalizes to more complex coding variants such as in-frame indels and stop-gains. Together, these results establish protein language models as an effective, accurate and general approach to predicting variant effects.

- Fowler and Fields (2014) — deep mutational scanning.

  Deep mutational scanning: a new style of protein science. Fowler & Fields (2014). *Nature Methods*. doi:10.1038/nmeth.3027

  > Mutagenesis provides insight into proteins, but only recently have assays that couple genotype to phenotype been used to assess the activities of as many as 1 million mutant versions of a protein in a single experiment. This approach-'deep mutational scanning'-yields large-scale data sets that can reveal intrinsic protein properties, protein behavior within cells and the consequences of human genetic variation. Deep mutational scanning is transforming the study of proteins, but many challenges must be tackled for it to fulfill its promise.

- Landrum et al. (2017), Chen et al. (2023) — ClinVar and gnomAD.

  ClinVar: improving access to variant interpretations and supporting evidence. Landrum et al. (2017). *Nucleic Acids Research*. doi:10.1093/nar/gkx1153

  > ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) is a freely available, public archive of human genetic variants and interpretations of their significance to disease, maintained at the National Institutes of Health. Interpretations of the clinical significance of variants are submitted by clinical testing laboratories, research laboratories, expert panels and other groups. ClinVar aggregates data by variant-disease pairs, and by variant (or set of variants). Data aggregated by variant are accessible on the website, in an improved set of variant call format files and as a new comprehensive XML report. ClinVar recently started accepting submissions that are focused primarily on providing phenotypic information for individuals who have had genetic testing. Submissions may come from clinical providers providing their own interpretation of the variant ('provider interpretation') or from groups such as patient registries that primarily provide phenotypic information from patients ('phenotyping only'). ClinVar continues to make improvements to

its search and retrieval functions. Several new fields are now indexed for more precise searching, and filters allow the user to narrow down a large set of search results.

A genomic mutational constraint map using variation in 76,156 human genomes. Chen et al. (2023). *Nature*. doi:10.1038/s41586-023-06045-0

> The depletion of disruptive variation caused by purifying natural selection (constraint) has been widely used to investigate protein-coding genes underlying human disorders 1-4 , but attempts to assess constraint for non-protein-coding regions have proved more difficult. Here we aggregate, process and release a dataset of 76,156 human genomes from the Genome Aggregation Database (gnomAD)-the largest public open-access human genome allele frequency reference dataset-and use it to build a genomic constraint map for the whole genome (genomic non-coding constraint of haploinsufficient variation (Gnocchi)). We present a refined mutational model that incorporates local sequence context and regional genomic features to detect depletions of variation. As expected, the average constraint for protein-coding sequences is stronger than that for non-coding regions. Within the non-coding genome, constrained regions are enriched for known regulatory elements and variants that are implicated in complex human diseases and traits, facilitating the triangulation of biological annotation, disease association and natural selection to non-coding DNA analysis. More constrained regulatory elements tend to regulate more constrained protein-coding genes, which in turn suggests that non-coding constraint can aid the identification of constrained genes that are as yet unrecognized by current gene constraint metrics. We demonstrate that this genome-wide constraint map improves the identification and interpretation of functional human genetic variation.

- Richards et al. (2015) — clinical guidelines for variant interpretation.

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Richards et al. (2015). *Genetics in Medicine*. doi:10.1038/gim.2015.30

> The American College of Medical Genetics and Genomics (ACMG) previously developed guidance for the interpretation of sequence variants.(1) In the past decade, sequencing technology has evolved rapidly with the advent of high-throughput next-generation sequencing. By adopting and leveraging next-generation sequencing, clinical laboratories are now performing an ever-increasing catalogue of genetic testing spanning genotyping, single genes, gene panels, exomes, genomes, transcriptomes, and epigenetic assays for genetic disorders. By virtue of increased complexity, this shift in genetic testing has been accompanied by new challenges in sequence interpretation. In this context the ACMG convened a workgroup in 2013 comprising representatives from the ACMG, the Association for Molecular Pathology (AMP), and the College of American Pathologists to revisit and revise the standards and guidelines for the interpretation of sequence variants. The group consisted of clinical laboratory directors and clinicians. This report represents expert opinion of the workgroup with input from ACMG, AMP, and College of American Pathologists stakeholders. These recommendations primarily apply to the breadth of genetic tests used in clinical laboratories, including genotyping, single genes, panels, exomes, and genomes. This report recommends the use of specific standard terminology-"pathogenic," "likely pathogenic," "uncertain significance," "likely benign," and "benign"-to describe variants identified in genes that cause Mendelian disorders. Moreover, this recommendation describes a process for classifying variants into these five categories based on criteria using typical types of variant evidence (e.g., population data, computational data, functional data, segregation data). Because of the increased complexity of analysis and interpretation of clinical genetic testing described in this report, the ACMG strongly recommends that clinical molecular genetic testing should be performed in a Clinical Laboratory Improvement Amendments-approved laboratory, with results interpreted by a board-certified clinical molecular geneticist or molecular genetic pathologist or the equivalent.

- The Critical Assessment of Genome Interpretation Consortium (2024) — CAGI.

CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. The Critical Assessment of Genome Interpretation Consortium (2024). *Genome Biology*. doi:10.1186/s13059-023-03113-6

Background The Critical Assessment of Genome Interpretation (CAGI) aims to advance the state-of-the-art for computational prediction of genetic variant impact, particularly where relevant to disease. The five complete editions of the CAGI community experiment comprised 50 challenges, in which participants made blind predictions of phenotypes from genetic data, and these were evaluated by independent assessors. Results Performance was particularly strong for clinical pathogenic variants, including some difficult-to-diagnose cases, and extends to interpretation of cancer-related variants. Missense variant interpretation methods were able to estimate biochemical effects with increasing accuracy. Assessment of methods for regulatory variants and complex trait disease risk was less definitive and indicates performance potentially suitable for auxiliary use in the clinic. Conclusions Results show that while current methods are imperfect, they have major utility for research and clinical applications. Emerging methods and increasingly large, robust datasets for training and assessment promise further progress ahead.

B.10.7 Benchmarking

- Radivojac et al. (2013), Jiang et al. (2016) — CAFA1 and CAFA2.

A large-scale evaluation of computational protein function prediction. Radivojac et al. (2013). *Nature Methods*. doi:10.1038/nmeth.2340

Automated annotation of protein function is challenging. As the number of sequenced genomes rapidly grows, the overwhelming majority of protein products can only be annotated computationally. If computational predictions are to be relied upon, it is crucial that the accuracy of these methods be high. Here we report the results from the first large-scale community-based critical assessment of protein function annotation (CAFA) experiment. Fifty-four methods representing the state of the art for protein function prediction were evaluated on a target set of 866 proteins from 11 organisms. Two findings stand out: (i) today's best protein function prediction algorithms substantially outperform widely used first-generation methods, with large gains on all types of targets; and (ii) although the top methods perform well enough to guide experiments, there is considerable need for improvement of currently available tools.

An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Jiang et al. (2016). *Genome Biology*. doi:10.1186/s13059-016-1037-6

Background A major bottleneck in our understanding of the molecular underpinnings of life is the assignment of function to proteins. While molecular experiments provide the most reliable annotation of proteins, their relatively low throughput and restricted purview have led to an increasing role for computational function prediction. However, assessing methods for protein function prediction and tracking progress in the field remain challenging. Results We conducted the second critical assessment of functional annotation (CAFA), a timed challenge to assess computational methods that automatically assign protein function. We evaluated 126 methods from 56 research groups for their ability to predict biological functions using Gene Ontology and gene-disease associations using Human Phenotype Ontology on a set of 3681 proteins from 18 species. CAFA2 featured expanded analysis compared with CAFA1, with regards to data set size, variety, and assessment metrics. To review progress in the field, the analysis compared the best methods from CAFA1 to those of CAFA2. Conclusions The top-performing methods in CAFA2 outperformed those from CAFA1. This increased accuracy can be attributed to a combination of the growing number of experimental annotations and improved methods for function prediction. The assessment also revealed that the definition of top-performing algorithms is ontology specific, that different performance metrics can be used to probe the nature of accurate predictions, and the relative diversity of predictions in the biological process and human phenotype ontologies. While there was methodological improvement between CAFA1 and CAFA2, the interpretation of results and usefulness of individual methods remain context-dependent.

- De Paolis Kaluza et al. (2026) — CAFA5.

Advances in Protein Function Prediction from the Fifth CAFA Challenge. De Paolis Kaluza et al. (2026). *bioRxiv (preprint)*. doi:10.64898/2026.04.27.716980

The Critical Assessment of Functional Annotation (CAFA) is a long-standing community effort to independently assess computational methods for protein function prediction, to highlight well-performing methodologies, to identify bottlenecks in the field, and to provide a forum for the dissemination of results and exchange of ideas. In its fifth round (CAFA5) of triennial challenges, a partnership with Kaggle Inc. facilitated participation from a large community of data scientists and computational biologists through a

> competitive prospective challenge on the crowdsourcing platform. In this work, we present an in-depth analysis of the submitted predictions and report improvements in accuracy over all methods from the previous CAFA challenges. We further introduce a new evaluation setting for proteins with pre-existing (incomplete) annotations and identify the need for methods that better leverage existing annotations to predict those that will be discovered later. Finally, we characterize the prospective evaluation framework by examining performance on a strict set of unpublished annotations and across intermediate database releases. Our results indicate that recent developments in the field, such as the availability of protein language models and accurately predicted 3D structures, as well as the growth of experimental annotations through biocuration, have all contributed to performance improvements. 1

- Clark and Radivojac (2013) — information-theoretic evaluation of ontology predictions.

  Information-theoretic evaluation of predicted ontological annotations. Clark & Radivojac (2013). *Bioinformatics*. doi:10.1093/bioinformatics/btt228

  > Motivation The development of effective methods for the prediction of ontological annotations is an important goal in computational biology, with protein function prediction and disease gene prioritization gaining wide recognition. Although various algorithms have been proposed for these tasks, evaluating their performance is difficult owing to problems caused both by the structure of biomedical ontologies and biased or incomplete experimental annotations of genes and gene products. Results We propose an information-theoretic framework to evaluate the performance of computational protein function prediction. We use a Bayesian network, structured according to the underlying ontology, to model the prior probability of a protein's function. We then define two concepts, misinformation and remaining uncertainty, that can be seen as information-theoretic analogs of precision and recall. Finally, we propose a single statistic, referred to as semantic distance, that can be used to rank classification models. We evaluate our approach by analyzing the performance of three protein function predictors of Gene Ontology terms and provide evidence that it addresses several weaknesses of currently used metrics. We believe this framework provides useful insights into the performance of protein function prediction tools. Supplementary information Supplementary data are available at Bioinformatics online.

B.11 Protein–protein interactions

Further papers for Chapter 11.

Papers marked ⋆ were mandatory in earlier reading lists. In the HTML version, click *Abstract* to show the abstract.

B.11.1 Experimental interactomes

- Fields and Song (1989) — the yeast two-hybrid system.

  A novel genetic system to detect protein–protein interactions. Fields & Song (1989). *Nature*. doi:10.1038/340245a0

  > Protein-protein interactions between two proteins have generally been studied using biochemical techniques such as crosslinking, co-immunoprecipitation and co-fractionation by chromatography. We have generated a novel genetic system to study these interactions by taking advantage of the properties of the GAL4 protein of the yeast Saccharomyces cerevisiae. This protein is a transcriptional activator required for the expression of genes encoding enzymes of galactose utilization. It consists of two separable and functionally essential domains: an N-terminal domain which binds to specific DNA sequences (UASG); and a C-terminal domain containing acidic regions, which is necessary to activate transcription. We have generated a system of two hybrid proteins containing parts of GAL4: the GAL4 DNA-binding domain fused to a protein 'X' and a GAL4 activating region fused to a protein 'Y'. If X and Y can form a protein-protein complex and reconstitute proximity of the GAL4 domains, transcription of a gene regulated by UASG occurs. We have tested this system using two yeast proteins that are known to interact--SNF1 and SNF4. High transcriptional activity is obtained only when both hybrids are present in a cell. This system may be applicable as a general method to identify proteins that interact with a known protein by the use of a simple galactose selection.

- Uetz et al. (2000), Ito et al. (2001) — two-hybrid screens of the yeast proteome.

  A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Uetz et al. (2000). *Nature*. doi:10.1038/35001009

  > Two large-scale yeast two-hybrid screens were undertaken to identify protein-protein interactions between full-length open reading frames predicted from the Saccharomyces cerevisiae genome sequence. In one approach, we constructed a protein array of about 6,000 yeast transformants, with each transformant expressing one of the open reading frames as a fusion to an activation domain. This array was screened by a simple and automated procedure for 192 yeast proteins, with positive responses identified by their positions in the array. In a second approach, we pooled cells expressing one of about 6,000 activation domain fusions to generate a library. We used a high-throughput screening procedure to screen nearly all of the 6,000 predicted yeast proteins, expressed as Gal4 DNA-binding domain fusion proteins, against the library, and characterized positives by sequence analysis. These approaches resulted in the detection of 957 putative interactions involving 1,004 S. cerevisiae proteins. These data reveal interactions that place functionally unclassified proteins in a biological context, interactions between proteins involved in the same biological function, and interactions that link biological functions together into larger cellular processes. The results of these screens are shown here.

  A comprehensive two-hybrid analysis to explore the yeast protein interactome. Ito et al. (2001). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.061034498

  > Protein-protein interactions play crucial roles in the execution of various biological functions. Accordingly, their comprehensive description would contribute considerably to the functional interpretation of fully sequenced genomes, which are flooded with novel genes of unpredictable functions. We previously developed a system to examine two-hybrid interactions in all possible combinations between the approximately 6,000 proteins of the budding yeast Saccharomyces cerevisiae. Here we have completed the comprehensive analysis using this system to identify 4,549 two-hybrid interactions among 3,278 proteins. Unexpectedly, these data do not largely overlap with those obtained by the other project [Uetz, P., et al. (2000) Nature (London) 403, 623-627] and hence have substantially expanded our knowledge on the protein interaction space or interactome of the yeast. Cumulative connection of these binary interactions generates a single huge network linking the vast majority of the proteins. Bioinformatics-aided selection of biologically relevant interactions highlights various intriguing subnetworks. They include, for instance, the one that had successfully foreseen the involvement of a novel protein in spindle pole body function as well as the one that may uncover a hitherto unidentified multiprotein complex potentially participating in the process of vesicular transport. Our data would thus significantly expand and improve the protein interaction map for the exploration of genome functions that eventually leads to thorough understanding of the cell as a molecular system.

- Rual et al. (2005), ⋆ Rolland et al. (2014) — towards the human binary interactome.

  Towards a proteome-scale map of the human protein–protein interaction network. Rual et al. (2005). *Nature*. doi:10.1038/nature04209

  > Systematic mapping of protein-protein interactions, or 'interactome' mapping, was initiated in model organisms, starting with defined biological processes and then expanding to the scale of the proteome. Although far from complete, such maps have revealed global topological and dynamic features of interactome networks that relate to known biological properties, suggesting that a human interactome map will provide insight into development and disease mechanisms at a systems level. Here we describe an initial version of a proteome-scale map of human binary protein-protein interactions. Using a stringent, high-throughput yeast two-hybrid system, we tested pairwise interactions among the products of approximately 8,100 currently available Gateway-cloned open reading frames and detected approximately 2,800 interactions. This data set, called CCSB-HI1, has a verification rate of approximately 78% as revealed by an independent co-affinity purification assay, and correlates significantly with other biological attributes. The CCSB-HI1 data set increases by approximately 70% the set of available binary interactions within the tested space and reveals more than 300 new connections to over 100 disease-associated proteins. This work represents an important step towards a systematic and comprehensive human interactome project.

  A Proteome-Scale Map of the Human Interactome Network. Rolland et al. (2014). *Cell*. doi:10.1016/j.cell.2014.10.050

> Just as reference genome sequences revolutionized human genetics, reference maps of interactome networks will be critical to fully understand genotype-phenotype relationships. Here, we describe a systematic map of ?14,000 high-quality human binary protein-protein interactions. At equal quality, this map is ?30% larger than what is available from small-scale studies published in the literature in the last few decades. While currently available information is highly biased and only covers a relatively small portion of the proteome, our systematic map appears strikingly more homogeneous, revealing a "broader" human interactome network than currently appreciated. The map also uncovers significant interconnectivity between known and candidate cancer gene products, providing unbiased evidence for an expanded functional cancer landscape, while demonstrating how high-quality interactome models will help "connect the dots" of the genomic revolution.

- Yu et al. (2008), Braun et al. (2008), Venkatesan et al. (2008) — high-quality binary maps and how to assess them.

High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Yu et al. (2008). *Science*. doi:10.1126/science.1158684

> Current yeast interactome network maps contain several hundred molecular complexes with limited and somewhat controversial representation of direct binary interactions. We carried out a comparative quality assessment of current yeast interactome data sets, demonstrating that high-throughput yeast two-hybrid (Y2H) screening provides high-quality binary interaction information. Because a large fraction of the yeast binary interactome remains to be mapped, we developed an empirically controlled mapping framework to produce a "second-generation" high-quality, high-throughput Y2H data set covering approximately 20% of all yeast binary interactions. Both Y2H and affinity purification followed by mass spectrometry (AP/MS) data are of equally high quality but of a fundamentally different and complementary nature, resulting in networks with different topological and biological properties. Compared to co-complex interactome models, this binary map is enriched for transient signaling interactions and intercomplex connections with a highly significant clustering between essential proteins. Rather than correlating with essentiality, protein connectivity correlates with genetic pleiotropy.

An experimentally derived confidence score for binary protein-protein interactions. Braun et al. (2008). *Nature Methods*. doi:10.1038/nmeth.1281

> Information on protein-protein interactions is of central importance for many areas of biomedical research. At present no method exists to systematically and experimentally assess the quality of individual interactions reported in interaction mapping experiments. To provide a standardized confidence-scoring method that can be applied to tens of thousands of protein interactions, we have developed an interaction tool kit consisting of four complementary, high-throughput protein interaction assays. We benchmarked these assays against positive and random reference sets consisting of well documented pairs of interacting human proteins and randomly chosen protein pairs, respectively. A logistic regression model was trained using the data from these reference sets to combine the assay outputs and calculate the probability that any newly identified interaction pair is a true biophysical interaction once it has been tested in the tool kit. This general approach will allow a systematic and empirical assignment of confidence scores to all individual protein-protein interactions in interactome networks.

An empirical framework for binary interactome mapping. Venkatesan et al. (2008). *Nature Methods*. doi:10.1038/nmeth.1280

> Several attempts have been made to systematically map protein-protein interaction, or 'interactome', networks. However, it remains difficult to assess the quality and coverage of existing data sets. Here we describe a framework that uses an empirically-based approach to rigorously dissect quality parameters of currently available human interactome maps. Our results indicate that high-throughput yeast two-hybrid (HT-Y2H) interactions for human proteins are more precise than literature-curated interactions supported by a single publication, suggesting that HT-Y2H is suitable to map a significant portion of the human interactome. We estimate that the human interactome contains approximately 130,000 binary interactions, most of which remain to be mapped. Similar to estimates of DNA sequence data quality and genome size early in the Human Genome Project, estimates of protein interaction data quality and interactome size are crucial to establish the magnitude of the task of comprehensive human interactome mapping and to elucidate a path toward this goal.

- Tarassov et al. (2008) — protein-fragment complementation in vivo.

An in Vivo Map of the Yeast Protein Interactome. Tarassov et al. (2008). *Science*. doi:10.1126/science.1153878

> Protein interactions regulate the systems-level behavior of cells; thus, deciphering the structure and dynamics of protein interaction networks in their cellular context is a central goal in biology. We have performed a genome-wide in vivo screen for protein-protein interactions in Saccharomyces cerevisiae by means of a protein-fragment complementation assay (PCA). We identified 2770 interactions among 1124 endogenously expressed proteins. Comparison with previous studies confirmed known interactions, but most were not known, revealing a previously unexplored subspace of the yeast protein interactome. The PCA detected structural and topological relationships between proteins, providing an 8-nanometer-resolution map of dynamically interacting complexes in vivo and extended networks that provide insights into fundamental cellular processes, including cell polarization and autophagy, pathways that are evolutionarily conserved and central to both development and human health.

- Rigaut et al. (1999) — tandem affinity purification.

A generic protein purification method for protein complex characterization and proteome exploration. Rigaut et al. (1999). *Nature Biotechnology*. doi:10.1038/13732

> We have developed a generic procedure to purify proteins expressed at their natural level under native conditions using a novel tandem affinity purification (TAP) tag. The TAP tag allows the rapid purification of complexes from a relatively small number of cells without prior knowledge of the complex composition, activity, or function. Combined with mass spectrometry, the TAP strategy allows for the identification of proteins interacting with a given target protein. The TAP method has been tested in yeast but should be applicable to other cells or organisms.

- Gavin et al. (2002), Ho et al. (2002), Krogan et al. (2006) — complexes of the yeast proteome.

Functional organization of the yeast proteome by systematic analysis of protein complexes. Gavin et al. (2002). *Nature*. doi:10.1038/415141a

> Most cellular processes are carried out by multiprotein complexes. The identification and analysis of their components provides insight into how the ensemble of expressed proteins (proteome) is organized into functional units. We used tandem-affinity purification (TAP) and mass spectrometry in a large-scale approach to characterize multiprotein complexes in Saccharomyces cerevisiae. We processed 1,739 genes, including 1,143 human orthologues of relevance to human biology, and purified 589 protein assemblies. Bioinformatic analysis of these assemblies defined 232 distinct multiprotein complexes and proposed new cellular roles for 344 proteins, including 231 proteins with no previous functional annotation. Comparison of yeast and human complexes showed that conservation across species extends from single proteins to their molecular environment. Our analysis provides an outline of the eukaryotic proteome as a network of protein complexes at a level of organization beyond binary interactions. This higher-order map contains fundamental biological information and offers the context for a more reasoned and informed approach to drug discovery.

Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Ho et al. (2002). *Nature*. doi:10.1038/415180a

> The recent abundance of genome sequence data has brought an urgent need for systematic proteomics to decipher the encoded protein networks that dictate cellular function. To date, generation of large-scale protein-protein interaction maps has relied on the yeast two-hybrid system, which detects binary interactions through activation of reporter gene expression. With the advent of ultrasensitive mass spectrometric protein identification methods, it is feasible to identify directly protein complexes on a proteome-wide scale. Here we report, using the budding yeast Saccharomyces cerevisiae as a test case, an example of this approach, which we term high-throughput mass spectrometric protein complex identification (HMS-PCI). Beginning with 10% of predicted yeast proteins as baits, we detected 3,617 associated proteins covering 25% of the yeast proteome. Numerous protein complexes were identified, including many new interactions in various signalling pathways and in the DNA damage response. Comparison of the HMS-PCI data set with interactions reported in the literature revealed an average threefold higher success rate in detection of known complexes compared with large-scale two-hybrid studies. Given the high degree of connectivity observed in this study, even partial HMS-PCI coverage of complex proteomes, including that of humans, should allow comprehensive identification of cellular networks.

Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Krogan et al. (2006). *Nature*. doi:10.1038/nature04670

> Identification of protein-protein interactions often provides insight into protein function, and many cellular processes are performed by stable protein complexes. We used tandem affinity purification to process 4,562 different tagged proteins of the yeast Saccharomyces cerevisiae. Each preparation was analysed by both matrix-assisted laser desorption/ionization-time of flight mass spectrometry and liquid chromatography tandem mass spectrometry to increase coverage and accuracy. Machine learning was used to integrate the mass spectrometry scores and assign probabilities to the protein-protein interactions. Among 4,087 different proteins identified with high confidence by mass spectrometry from 2,357 successful purifications, our core data set (median precision of 0.69) comprises 7,123 protein-protein interactions involving 2,708 proteins. A Markov clustering algorithm organized these interactions into 547 protein complexes averaging 4.9 subunits per complex, about half of them absent from the MIPS database, as well as 429 additional interactions between pairs of complexes. The data (all of which are available online) will help future studies on individual proteins as well as functional genomics and systems biology.

- Huttlin et al. (2015), Huttlin et al. (2021) — BioPlex.

The BioPlex Network: A Systematic Exploration of the Human Interactome. Huttlin et al. (2015). *Cell*. doi:10.1016/j.cell.2015.06.043

> Protein interactions form a network whose structure drives cellular function and whose organization informs biological inquiry. Using high-throughput affinity-purification mass spectrometry, we identify interacting partners for 2,594 human proteins in HEK293T cells. The resulting network (BioPlex) contains 23,744 interactions among 7,668 proteins with 86% previously undocumented. BioPlex accurately depicts known complexes, attaining 80%-100% coverage for most CORUM complexes. The network readily subdivides into communities that correspond to complexes or clusters of functionally related proteins. More generally, network architecture reflects cellular localization, biological process, and molecular function, enabling functional characterization of thousands of proteins. Network structure also reveals associations among thousands of protein domains, suggesting a basis for examining structurally related proteins. Finally, BioPlex, in combination with other approaches, can be used to reveal interactions of biological or clinical significance. For example, mutations in the membrane protein VAPB implicated in familial amyotrophic lateral sclerosis perturb a defined community of interactors.

Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Huttlin et al. (2021). *Cell*. doi:10.1016/j.cell.2021.04.011

> Thousands of interactions assemble proteins into modules that impart spatial and functional organization to the cellular proteome. Through affinity-purification mass spectrometry, we have created two proteome-scale, cell-line-specific interaction networks. The first, BioPlex 3.0, results from affinity purification of 10,128 human proteins-half the proteome-in 293T cells and includes 118,162 interactions among 14,586 proteins. The second results from 5,522 immunoprecipitations in HCT116 cells. These networks model the interactome whose structure encodes protein function, localization, and complex membership. Comparison across cell lines validates thousands of interactions and reveals extensive customization. Whereas shared interactions reside in core complexes and involve essential proteins, cell-specific interactions link these complexes, "rewiring" subnetworks within each cell's interactome. Interactions covary among proteins of shared function as the proteome remodels to produce each cell's phenotype. Viewable interactively online through BioPlexExplorer, these networks define principles of proteome organization and enable unknown protein characterization.

- Drew et al. (2017), Drew et al. (2021), Fischer et al. (2025) — hu.MAP.

Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Drew et al. (2017). *Molecular Systems Biology*. doi:10.15252/msb.20167490

> Macromolecular protein complexes carry out many of the essential functions of cells, and many genetic diseases arise from disrupting the functions of such complexes. Currently, there is great interest in defining the complete set of human protein complexes, but recent published maps lack comprehensive coverage. Here, through the synthesis of over 9,000 published mass spectrometry experiments, we present hu.MAP, the most comprehensive and accurate human protein complex map to date, containing > 4,600 total complexes, > 7,700 proteins, and > 56,000 unique interactions, including thousands of confident protein interactions not identified by the original publications. hu.MAP accurately recapitulates known complexes withheld from the learning procedure, which was optimized with the aid of a new quantitative metric (k

-cliques) for comparing sets of sets. The vast majority of complexes in our map are significantly enriched with literature annotations, and the map overall shows improved coverage of many disease-associated proteins, as we describe in detail for ciliopathies. Using hu.MAP, we predicted and experimentally validated candidate ciliopathy disease genes in vivo in a model vertebrate, discovering CCDC138, WDR90, and KIAA1328 to be new cilia basal body/centriolar satellite proteins, and identifying ANKRD55 as a novel member of the intraflagellar transport machinery. By offering significant improvements to the accuracy and coverage of human protein complexes, hu.MAP (http://proteincomplexes.org) serves as a valuable resource for better understanding the core cellular functions of human proteins and helping to determine mechanistic foundations of human disease.

hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Drew et al. (2021). *Molecular Systems Biology*. doi:10.15252/msb.202010016

> A general principle of biology is the self-assembly of proteins into functional complexes. Characterizing their composition is, therefore, required for our understanding of cellular functions. Unfortunately, we lack knowledge of the comprehensive set of identities of protein complexes in human cells. To address this gap, we developed a machine learning framework to identify protein complexes in over 15,000 mass spectrometry experiments which resulted in the identification of nearly 7,000 physical assemblies. We show our resource, hu.MAP 2.0, is more accurate and comprehensive than previous state of the art high-throughput protein complex resources and gives rise to many new hypotheses, including for 274 completely uncharacterized proteins. Further, we identify 253 promiscuous proteins that participate in multiple complexes pointing to possible moonlighting roles. We have made hu.MAP 2.0 easily searchable in a web interface (http://humap2.proteincomplexes.org/), which will be a valuable resource for researchers across a broad range of interests including systems biology, structural biology, and molecular explanations of disease.

hu.MAP3.0: atlas of human protein complexes by integration of >25,000 proteomic experiments. Fischer et al. (2025). *Molecular Systems Biology*. doi:10.1038/s44320-025-00121-5

> Macromolecular protein complexes carry out most cellular functions. Unfortunately, we lack the subunit composition for many human protein complexes. To address this gap we integrated >25,000 mass spectrometry experiments using a machine learning approach to identify >15,000 human protein complexes. We show our map of protein complexes is highly accurate and more comprehensive than previous maps, placing nearly 70% of human proteins into their physical contexts. We globally characterize our complexes using mass spectrometry based protein covariation data (ProteomeHD.2) and identify covarying complexes suggesting common functional associations. hu.MAP3.0 generates testable functional hypotheses for 472 uncharacterized proteins which we support using AlphaFold modeling. Additionally, we use AlphaFold modeling to identify 5871 mutually exclusive proteins in hu.MAP3.0 complexes suggesting complexes serve different functional roles depending on their subunit composition. We identify expression as the primary way cells and organisms relieve the conflict of mutually exclusive subunits. Finally, we import our complexes to EMBL-EBI's Complex Portal (https://www.ebi.ac.uk/complexportal/home) and provide complexes through our hu.MAP3.0 web interface (https://humap3.proteincomplexes.org/). We expect our resource to be highly impactful to the broader research community.

- Costanzo et al. (2016) — the yeast genetic interaction network.

A global genetic interaction network maps a wiring diagram of cellular function. Costanzo et al. (2016). *Science*. doi:10.1126/science.aaf1420

> We generated a global genetic interaction network for Saccharomyces cerevisiae, constructing more than 23 million double mutants, identifying about 550,000 negative and about 350,000 positive genetic interactions. This comprehensive network maps genetic interactions for essential gene pairs, highlighting essential genes as densely connected hubs. Genetic interaction profiles enabled assembly of a hierarchical model of cell function, including modules corresponding to protein complexes and pathways, biological processes, and cellular compartments. Negative interactions connected functionally related genes, mapped core bioprocesses, and identified pleiotropic genes, whereas positive interactions often mapped general regulatory connections among gene pairs, rather than shared functionality. The global network illustrates how coherent sets of genetic interactions connect protein complex and pathway modules to map a functional wiring diagram of the cell.

B.11.2 Regulatory networks

- Ideker et al. (2000), Ideker et al. (2001) — perturbations and integrated analysis.

  Discovery of regulatory interactions through perturbation: inference and experimental design. Ideker et al. (2000). *Pacific Symposium on Biocomputing*.

  Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Ideker et al. (2001). *Science*. doi:10.1126/science.292.5518.929

  > We demonstrate an integrated approach to build, test, and refine a model of a cellular pathway, in which perturbations to critical pathway components are analyzed using DNA microarrays, quantitative proteomics, and databases of known physical interactions. Using this approach, we identify 997 messenger RNAs responding to 20 systematic perturbations of the yeast galactose-utilization pathway, provide evidence that approximately 15 of 289 detected proteins are regulated posttranscriptionally, and identify explicit physical interactions governing the cellular response to each perturbation. We refine the model through further iterations of perturbation and global measurements, suggesting hypotheses about the regulation of galactose utilization and physical interactions between this and a variety of other metabolic pathways.

- Gardner et al. (2003), Tegnér et al. (2003) — inferring networks from perturbations.

  Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Gardner et al. (2003). *Science*. doi:10.1126/science.1081900

  > The complexity of cellular gene, protein, and metabolite networks can hinder attempts to elucidate their structure and function. To address this problem, we used systematic transcriptional perturbations to construct a first-order model of regulatory interactions in a nine-gene subnetwork of the SOS pathway in Escherichia coli. The model correctly identified the major regulatory genes and the transcriptional targets of mitomycin C activity in the subnetwork. This approach, which is experimentally and computationally scalable, provides a framework for elucidating the functional properties of genetic networks and identifying molecular targets of pharmacological compounds.

  Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. Tegnér et al. (2003). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0933416100

  > While the fundamental building blocks of biology are being tabulated by the various genome projects, microarray technology is setting the stage for the task of deducing the connectivity of large-scale gene networks. We show how the perturbation of carefully chosen genes in a microarray experiment can be used in conjunction with a reverse engineering algorithm to reveal the architecture of an underlying gene regulatory network. Our iterative scheme identifies the network topology by analyzing the steady-state changes in gene expression resulting from the systematic perturbation of a particular node in the network. We highlight the validity of our reverse engineering approach through the successful deduction of the topology of a linear in numero gene network and a recently reported model for the segmentation polarity network in Drosophila melanogaster. Our method may prove useful in identifying and validating specific drug targets and in deconvolving the effects of chemical compounds.

- Margolin et al. (2006), Huynh-Thu et al. (2010) — ARACNE and GENIE3.

  ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. Margolin et al. (2006). *BMC Bioinformatics*. doi:10.1186/1471-2105-7-s1-s7

  > Background Elucidating gene regulatory networks is crucial for understanding normal cell physiology and complex pathologic phenotypes. Existing computational methods for the genome-wide "reverse engineering" of such networks have been successful only for lower eukaryotes with simple genomes. Here we present ARACNE, a novel algorithm, using microarray expression profiles, specifically designed to scale up to the complexity of regulatory networks in mammalian cells, yet general enough to address a wider range of network deconvolution problems. This method uses an information theoretic approach to eliminate the majority of indirect interactions inferred by co-expression methods. Results We prove that ARACNE reconstructs the network exactly (asymptotically) if the effect of loops in the network topology is

> negligible, and we show that the algorithm works well in practice, even in the presence of numerous loops and complex topologies. We assess ARACNE's ability to reconstruct transcriptional regulatory networks using both a realistic synthetic dataset and a microarray dataset from human B cells. On synthetic datasets ARACNE achieves very low error rates and outperforms established methods, such as Relevance Networks and Bayesian Networks. Application to the deconvolution of genetic networks in human B cells demonstrates ARACNE's ability to infer validated transcriptional targets of the cMYC proto-oncogene. We also study the effects of misestimation of mutual information on network reconstruction, and show that algorithms based on mutual information ranking are more resilient to estimation errors. Conclusion ARACNE shows promise in identifying direct transcriptional interactions in mammalian cellular networks, a problem that has challenged existing reverse engineering algorithms. This approach should enhance our ability to use microarray data to elucidate functional mechanisms that underlie cellular processes and to identify molecular targets of pharmacological compounds in mammalian cellular networks.

Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. Huynh-Thu et al. (2010). *PLoS ONE*. doi:10.1371/journal.pone.0012776

> One of the pressing open problems of computational systems biology is the elucidation of the topology of genetic regulatory networks (GRNs) using high throughput genomic data, in particular microarray gene expression data. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge aims to evaluate the success of GRN inference algorithms on benchmarks of simulated data. In this article, we present GENIE3, a new algorithm for the inference of GRNs that was best performer in the DREAM4 In Silico Multifactorial challenge. GENIE3 decomposes the prediction of a regulatory network between p genes into p different regression problems. In each of the regression problems, the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes), using tree-based ensemble methods Random Forests or Extra-Trees. The importance of an input gene in the prediction of the target gene expression pattern is taken as an indication of a putative regulatory link. Putative regulatory links are then aggregated over all genes to provide a ranking of interactions from which the whole network is reconstructed. In addition to performing well on the DREAM4 In Silico Multifactorial challenge simulated data, we show that GENIE3 compares favorably with existing algorithms to decipher the genetic regulatory network of Escherichia coli. It doesn't make any assumption about the nature of gene regulation, can deal with combinatorial and non-linear interactions, produces directed GRNs, and is fast and scalable. In conclusion, we propose a new algorithm for GRN inference that performs well on both synthetic and real gene expression data. The algorithm, based on feature selection with tree-based ensemble methods, is simple and generic, making it adaptable to other types of genomic data and interactions.

- Marbach et al. (2012) — DREAM5: the wisdom of crowds.

Wisdom of crowds for robust gene network inference. Marbach et al. (2012). *Nature Methods*. doi:10.1038/nmeth.2016

> Reconstructing gene regulatory networks from high-throughput data is a long-standing challenge. Through the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, we performed a comprehensive blind assessment of over 30 network inference methods on Escherichia coli, Staphylococcus aureus, Saccharomyces cerevisiae and in silico microarray data. We characterize the performance, data requirements and inherent biases of different inference approaches, and we provide guidelines for algorithm application and development. We observed that no single inference method performs optimally across all data sets. In contrast, integration of predictions from multiple inference methods shows robust and high performance across diverse data sets. We thereby constructed high-confidence networks for E. coli and S. aureus, each comprising ~1,700 transcriptional interactions at a precision of ~50%. We experimentally tested 53 previously unobserved regulatory interactions in E. coli, of which 23 (43%) were supported. Our results establish community-based methods as a powerful and robust tool for the inference of transcriptional gene regulatory networks.

- Pratapa et al. (2020) — benchmarking network inference on single-cell data.

Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Pratapa et al. (2020). *Nature Methods*. doi:10.1038/s41592-019-0690-6

> We present a systematic evaluation of state-of-the-art algorithms for inferring gene regulatory networks from single-cell transcriptional data. As the ground truth for assessing accuracy, we use synthetic networks with predictable trajectories, literature-curated Boolean models and diverse transcriptional regulatory networks. We develop a strategy to simulate single-cell transcriptional data from synthetic

and Boolean networks that avoids pitfalls of previously used methods. Furthermore, we collect networks from multiple experimental single-cell RNA-seq datasets. We develop an evaluation framework called BEELINE. We find that the area under the precision-recall curve and early precision of the algorithms are moderate. The methods are better in recovering interactions in synthetic networks than Boolean models. The algorithms with the best early precision values for Boolean models also perform well on experimental datasets. Techniques that do not require pseudotime-ordered cells are generally more accurate. Based on these results, we present recommendations to end users. BEELINE will aid the development of gene regulatory network inference algorithms.

- Dixit et al. (2016), Replogle et al. (2022) — Perturb-seq.

  Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Dixit et al. (2016). *Cell*. doi:10.1016/j.cell.2016.11.038

  > Genetic screens help infer gene function in mammalian cells, but it has remained difficult to assay complex phenotypes-such as transcriptional profiles-at scale. Here, we develop Perturb-seq, combining single-cell RNA sequencing (RNA-seq) and clustered regularly interspaced short palindromic repeats (CRISPR)-based perturbations to perform many such assays in a pool. We demonstrate Perturb-seq by analyzing 200,000 cells in immune cells and cell lines, focusing on transcription factors regulating the response of dendritic cells to lipopolysaccharide (LPS). Perturb-seq accurately identifies individual gene targets, gene signatures, and cell states affected by individual perturbations and their genetic interactions. We posit new functions for regulators of differentiation, the anti-viral response, and mitochondrial function during immune activation. By decomposing many high content measurements into the effects of perturbations, their interactions, and diverse cell metadata, Perturb-seq dramatically increases the scope of pooled genomic assays.

  Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Replogle et al. (2022). *Cell*. doi:10.1016/j.cell.2022.05.013

  > A central goal of genetics is to define the relationships between genotypes and phenotypes. High-content phenotypic screens such as Perturb-seq (CRISPR-based screens with single-cell RNA-sequencing readouts) enable massively parallel functional genomic mapping but, to date, have been used at limited scales. Here, we perform genome-scale Perturb-seq targeting all expressed genes with CRISPR interference (CRISPRi) across >2.5 million human cells. We use transcriptional phenotypes to predict the function of poorly characterized genes, uncovering new regulators of ribosome biogenesis (including CCDC86, ZNF236, and SPATA5L1), transcription (C7orf26), and mitochondrial respiration (TMEM242). In addition to assigning gene function, single-cell transcriptional phenotypes allow for in-depth dissection of complex cellular phenomena-from RNA processing to differentiation. We leverage this ability to systematically identify genetic drivers and consequences of aneuploidy and to discover an unanticipated layer of stress-specific regulation of the mitochondrial genome. Our information-rich genotype-phenotype map reveals a multidimensional portrait of gene and cellular function.

B.11.3 Network concepts

- Barabási and Albert (1999), Jeong et al. (2000) — scale-free networks.

  Emergence of Scaling in Random Networks. Barabási & Albert (1999). *Science*. doi:10.1126/science.286.5439.509

  > Systems as diverse as genetic networks or the World Wide Web are best described as networks with complex topology. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

  The large-scale organization of metabolic networks. Jeong et al. (2000). *Nature*. doi:10.1038/35036627

> In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. We show that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems. This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks, and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents.

- Watts and Strogatz (1998) — small-world networks.

  Collective dynamics of 'small-world' networks. Watts & Strogatz (1998). *Nature*. doi:10.1038/30918

  > Networks of coupled dynamical systems have been used to model biological oscillators, Josephson junction arrays, excitable media, neural networks, spatial games, genetic control networks and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes. Here we explore simple models of networks that can be tuned through this middle ground: regular networks 'rewired' to introduce increasing amounts of disorder. We find that these systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. We call them 'small-world' networks, by analogy with the small-world phenomenon (popularly known as six degrees of separation. The neural network of the worm Caenorhabditis elegans, the power grid of the western United States, and the collaboration graph of film actors are shown to be small-world networks. Models of dynamical systems with small-world coupling display enhanced signal-propagation speed, computational power, and synchronizability. In particular, infectious diseases spread more easily in small-world networks than in regular lattices.

- Jeong et al. (2001), Han et al. (2004), Ekman et al. (2006) — hubs.

  Lethality and centrality in protein networks. Jeong et al. (2001). *Nature*. doi:10.1038/35075138

  Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Han et al. (2004). *Nature*. doi:10.1038/nature02555

  > In apparently scale-free protein-protein interaction networks, or 'interactome' networks, most proteins interact with few partners, whereas a small but significant proportion of proteins, the 'hubs', interact with many partners. Both biological and non-biological scale-free networks are particularly resistant to random node removal but are extremely sensitive to the targeted removal of hubs. A link between the potential scale-free topology of interactome networks and genetic robustness seems to exist, because knockouts of yeast genes encoding hubs are approximately threefold more likely to confer lethality than those of non-hubs. Here we investigate how hubs might contribute to robustness and other cellular properties for protein-protein interactions dynamically regulated both in time and in space. We uncovered two types of hub: 'party' hubs, which interact with most of their partners simultaneously, and 'date' hubs, which bind their different partners at different times or locations. Both in silico studies of network connectivity and genetic interactions described in vivo support a model of organized modularity in which date hubs organize the proteome, connecting biological processes--or modules--to each other, whereas party hubs function inside modules.

  What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?. Ekman et al. (2006). *Genome Biology*. doi:10.1186/gb-2006-7-6-r45

  > Background Most proteins interact with only a few other proteins while a small number of proteins (hubs) have many interaction partners. Hub proteins and non-hub proteins differ in several respects; however, understanding is not complete about what properties characterize the hubs and set them apart from proteins of low connectivity. Therefore, we have investigated what differentiates hubs from non-hubs and static hubs (party hubs) from dynamic hubs (date hubs) in the protein-protein interaction network of Saccharomyces cerevisiae. Results The many interactions of hub proteins can only partly be explained by bindings to similar proteins or domains. It is evident that domain repeats, which are associated with binding, are enriched in hubs. Moreover, there is an over representation of multi-domain proteins and

long proteins among the hubs. In addition, there are clear differences between party hubs and date hubs. Fewer of the party hubs contain long disordered regions compared to date hubs, indicating that these regions are important for flexible binding but less so for static interactions. Furthermore, party hubs interact to a large extent with each other, supporting the idea of party hubs as the cores of highly clustered functional modules. In addition, hub proteins, and in particular party hubs, are more often ancient. Finally, the more recent paralogs of party hubs are underrepresented. Conclusion Our results indicate that multiple and repeated domains are enriched in hub proteins and, further, that long disordered regions, which are common in date hubs, are particularly important for flexible binding.

- Maslov and Sneppen (2002) — degree correlations in protein networks.

  Specificity and Stability in Topology of Protein Networks. Maslov & Sneppen (2002). *Science*. doi:10.1126/science.1065103

  > Molecular networks guide the biochemistry of a living cell on multiple levels: Its metabolic and signaling pathways are shaped by the network of interacting proteins, whose production, in turn, is controlled by the genetic regulatory network. To address topological properties of these two networks, we quantified correlations between connectivities of interacting nodes and compared them to a null model of a network, in which all links were randomly rewired. We found that for both interaction and regulatory networks, links between highly connected proteins are systematically suppressed, whereas those between a highly connected and low-connected pairs of proteins are favored. This effect decreases the likelihood of cross talk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of deleterious perturbations.

- Milo et al. (2002), Ravasz et al. (2002) — motifs and hierarchical modularity.

  Network Motifs: Simple Building Blocks of Complex Networks. Milo et al. (2002). *Science*. doi:10.1126/science.298.5594.824

  > Complex networks are studied across many fields of science. To uncover their structural design principles, we defined "network motifs," patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. We found such motifs in networks from biochemistry, neurobiology, ecology, and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of Escherichia coli and Saccharomyces cerevisiae or from those found in the World Wide Web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in Caenorhabditis elegans. Motifs may thus define universal classes of networks. This approach may uncover the basic building blocks of most networks.

  Hierarchical Organization of Modularity in Metabolic Networks. Ravasz et al. (2002). *Science*. doi:10.1126/science.1073374

  > Spatially or chemically isolated functional modules composed of several cellular components and carrying discrete functions are considered fundamental building blocks of cellular organization, but their presence in highly integrated biochemical networks lacks quantitative support. Here, we show that the metabolic networks of 43 distinct organisms are organized into many small, highly connected topologic modules that combine in a hierarchical manner into larger, less cohesive units, with their number and degree of clustering following a power law. Within Escherichia coli, the uncovered hierarchical modularity closely overlaps with known metabolic functions. The identified network architecture may be generic to system-level cellular organization.

- Girvan and Newman (2002), Enright (2002) — community detection and MCL.

  Community structure in social and biological networks. Girvan & Newman (2002). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.122653799

  > A number of recent studies have focused on the statistical properties of networked systems such as social networks and the Worldwide Web. Researchers have concentrated particularly on a few properties that seem to be common to many networks: the small-world property, power-law degree distributions, and network transitivity. In this article, we highlight another property that is found in many networks, the property of community structure, in which network nodes are joined together in tightly knit groups, between which there are only looser connections. We propose a method for detecting such communities, built around the idea of using centrality indices to find community boundaries. We test our method on

computer-generated and real-world graphs whose community structure is already known and find that the method detects this known structure with high sensitivity and reliability. We also apply the method to two networks whose community structure is not well known--a collaboration network and a food web--and find that it detects significant and informative community divisions in both cases.

An efficient algorithm for large-scale detection of protein families. Enright (2002). *Nucleic Acids Research*. doi:10.1093/nar/30.7.1575

> Detection of protein families in large databases is one of the principal research objectives in structural and functional genomics. Protein family classification can significantly contribute to the delineation of functional diversity of homologous proteins, the prediction of function based on domain architecture or the presence of sequence motifs as well as comparative genomics, providing valuable evolutionary insights. We present a novel approach called TRIBE-MCL for rapid and accurate clustering of protein sequences into families. The method relies on the Markov cluster (MCL) algorithm for the assignment of proteins into families based on precomputed sequence similarity information. This novel approach does not suffer from the problems that normally hinder other protein sequence clustering algorithms, such as the presence of multi-domain proteins, promiscuous domains and fragmented proteins. The method has been rigorously tested and validated on a number of very large databases, including SwissProt, InterPro, SCOP and the draft human genome. Our results indicate that the method is ideally suited to the rapid and accurate detection of protein families on a large scale. The method has been used to detect and categorise protein families within the draft human genome and the resulting families have been used to annotate a large proportion of human proteins.

- Hakes et al. (2008), Lima-Mendez and Helden (2009), Broido and Clauset (2019) — critical views of network biology.

Protein-protein interaction networks and biology—what's the connection?. Hakes et al. (2008). *Nature Biotechnology*. doi:10.1038/nbt0108-69

The powerful law of the power law and other myths in network biology. Lima-Mendez & van Helden (2009). *Molecular BioSystems*. doi:10.1039/b908681a

> For almost 10 years, topological analysis of different large-scale biological networks (metabolic reactions, protein interactions, transcriptional regulation) has been highlighting some recurrent properties: power law distribution of degree, scale-freeness, small world, which have been proposed to confer functional advantages such as robustness to environmental changes and tolerance to random mutations. Stochastic generative models inspired different scenarios to explain the growth of interaction networks during evolution. The power law and the associated properties appeared so ubiquitous in complex networks that they were qualified as "universal laws". However, these properties are no longer observed when the data are subjected to statistical tests: in most cases, the data do not fit the expected theoretical models, and the cases of good fitting merely result from sampling artefacts or improper data representation. The field of network biology seems to be founded on a series of myths, i.e. widely believed but false ideas. The weaknesses of these foundations should however not be considered as a failure for the entire domain. Network analysis provides a powerful frame for understanding the function and evolution of biological processes, provided it is brought to an appropriate level of description, by focussing on smaller functional modules and establishing the link between their topological properties and their dynamical behaviour.

Scale-free networks are rare. Broido & Clauset (2019). *Nature Communications*. doi:10.1038/s41467-019-08746-5

> Real-world networks are often claimed to be scale free, meaning that the fraction of nodes with degree k follows a power law k -α , a pattern with broad implications for the structure and dynamics of complex systems. However, the universality of scale-free networks remains controversial. Here, we organize different definitions of scale-free networks and construct a severe test of their empirical prevalence using state-of-the-art statistical tools applied to nearly 1000 social, biological, technological, transportation, and information networks. Across these networks, we find robust evidence that strongly scale-free structure is empirically rare, while for most networks, log-normal distributions fit the data as well or better than power laws. Furthermore, social networks are at best weakly scale free, while a handful of technological and biological networks appear strongly scale free. These findings highlight the structural diversity of real-world networks and the need for new theoretical explanations of these non-scale-free patterns.

- Yanai and DeLisi (2002), Rhodes et al. (2005) — networks from comparative genomics and probabilistic integration.

  The society of genes: networks of functional links between genes from comparative genomics. Yanai & DeLisi (2002). *Genome Biology*. doi:10.1186/gb-2002-3-11-research0064

  > Background Comparative genomics provides at least three methods beyond traditional sequence similarity for identifying functional links between genes: the examination of common phylogenetic distributions, the analysis of conserved proximity along the chromosomes of multiple genomes, and observations of fusions of genes into a multidomain gene in another organism. We have previously generated the links according to each of these methods individually for 43 known microbial genomes. Here we combine these results to construct networks of functional associations. Results We show that the functional networks obtained by applying these methods have different topologies and that the information they provide is largely additive. In particular, the combined networks of functional links contain an average of 57% of an organism's complete genetic complement, uncover substantial portions of known pathways, and suggest the function of previously unannotated genes. In addition, the combined networks are qualitatively different from the networks obtained using individual methods. They have a dominant cluster that contains approximately 80%-90% of the genes, independent of genome size, and the dominant clusters show the small world behavior expected of a biological system, with global connectivity that is nearly random, and local properties that are highly ordered. Conclusions When the information on functional linkage provided by three emerging computational methods is combined, the integrated network uncovers large numbers of conserved pathways and identifies clusters of functionally related genes. It therefore shows considerable utility and promise as a tool for understanding genomic structure, and for guiding high throughput experimental investigations.

  Probabilistic model of the human protein-protein interaction network. Rhodes et al. (2005). *Nature Biotechnology*. doi:10.1038/nbt1103

  > A catalog of all human protein-protein interactions would provide scientists with a framework to study protein deregulation in complex diseases such as cancer. Here we demonstrate that a probabilistic analysis integrating model organism interactome data, protein domain data, genome-wide gene expression data and functional annotation data predicts nearly 40,000 protein-protein interactions in humans-a result comparable to those obtained with experimental and computational approaches in model organisms. We validated the accuracy of the predictive model on an independent test set of known interactions and also experimentally confirmed two predicted interactions relevant to human cancer, implicating uncharacterized proteins into definitive pathways. We also applied the human interactome network to cancer genomics data and identified several interaction subnetworks activated in cancer. This integrative analysis provides a comprehensive framework for exploring the human protein interaction network.

B.11.4 Functional association networks

- ⋆ Guala et al. (2019) — genome-wide functional association networks.

  Genome-wide functional association networks: background, data & state-of-the-art resources. Guala et al. (2019). *Briefings in Bioinformatics*. doi:10.1093/bib/bbz064

  > The vast amount of experimental data from recent advances in the field of high-throughput biology begs for integration into more complex data structures such as genome-wide functional association networks. Such networks have been used for elucidation of the interplay of intra-cellular molecules to make advances ranging from the basic science understanding of evolutionary processes to the more translational field of precision medicine. The allure of the field has resulted in rapid growth of the number of available network resources, each with unique attributes exploitable to answer different biological questions. Unfortunately, the high volume of network resources makes it impossible for the intended user to select an appropriate tool for their particular research question. The aim of this paper is to provide an overview of the underlying data and representative network resources as well as to mention methods of integration, allowing a customized approach to resource selection. Additionally, this report will provide a primer for researchers venturing into the field of network integration.

- Jansen et al. (2003), Lee et al. (2004) — Bayesian integration of evidence.

A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Jansen et al. (2003). *Science*. doi:10.1126/science.1087361

> We have developed an approach using Bayesian networks to predict protein-protein interactions genome-wide in yeast. Our method naturally weights and combines into reliable predictions genomic features only weakly associated with interaction (e.g., messenger RNAcoexpression, coessentiality, and colocalization). In addition to de novo predictions, it can integrate often noisy, experimental interaction data sets. We observe that at given levels of sensitivity, our predictions are more accurate than the existing high-throughput experimental data sets. We validate our predictions with TAP (tandem affinity purification) tagging experiments. Our analysis, which gives a comprehensive view of yeast interactions, is available at genecensus.org/intint.

A Probabilistic Functional Network of Yeast Genes. Lee et al. (2004). *Science*. doi:10.1126/science.1099511

> A conceptual framework for integrating diverse functional genomics data was developed by reinterpreting experiments to provide numerical likelihoods that genes are functionally linked. This allows direct comparison and integration of different classes of data. The resulting probabilistic gene network estimates the functional coupling between genes. Within this framework, we reconstructed an extensive, high-quality functional gene network for Saccharomyces cerevisiae, consisting of 4681 (approximately 81%) of the known yeast genes linked by approximately 34,000 probabilistic linkages comparable in accuracy to small-scale interaction assays. The integrated linkages distinguish true from false-positive interactions in earlier data sets; new interactions emerge from genes' network contexts, as shown for genes in chromatin modification and ribosome biogenesis.

- Alexeyenko and Sonnhammer (2009), Buzzao, Persson, et al. (2024) — FunCoup.

Global networks of functional coupling in eukaryotes from comprehensive data integration. Alexeyenko & Sonnhammer (2009). *Genome Research*. doi:10.1101/gr.087528.108

> No single experimental method can discover all connections in the interactome. A computational approach can help by integrating data from multiple, often unrelated, proteomics and genomics pipelines. Reconstructing global networks of functional coupling (FC) faces the challenges of scale and heterogeneity--how to efficiently integrate huge amounts of diverse data from multiple organisms, yet ensuring high accuracy. We developed FunCoup, an optimized Bayesian framework, to resolve these issues. Because interactomes comprise functional coupling of many types, FunCoup annotates network edges with confidence scores in support of different kinds of interactions: physical interaction, protein complex member, metabolic, or signaling link. This capability boosted overall accuracy. On the whole, the constructed framework was comprehensively tested to optimize the overall confidence and ensure seamless, automated incorporation of new data sets of heterogeneous types. Using over 50 data sets in seven organisms and extensively transferring information between orthologs, FunCoup predicted global networks in eight eukaryotes. For the Ciona intestinalis network, only orthologous information was used, and it recovered a significant number of experimental facts. FunCoup predictions were validated on independent cancer mutation data. We show how FunCoup can be used for discovering candidate members of the Parkinson and Alzheimer pathways. Cross-species pathway conservation analysis provided further support to these observations.

FunCoup 6: advancing functional association networks across species with directed links and improved user experience. Buzzao et al. (2024). *Nucleic Acids Research*. doi:10.1093/nar/gkae1021

> FunCoup 6 (https://funcoup.org) represents a significant advancement in global functional association networks, aiming to provide researchers with a comprehensive view of the functional coupling interactome. This update introduces novel methodologies and integrated tools for improved network inference and analysis. Major new developments in FunCoup 6 include vastly expanding the coverage of gene regulatory links, a new framework for bin-free Bayesian training and a new website. FunCoup 6 integrates a new tool for disease and drug target module identification using the TOPAS algorithm. To expand the utility of the resource for biomedical research, it incorporates pathway enrichment analysis using the ANUBIX and EASE algorithms. The unique comparative interactomics analysis in FunCoup provides insights of network conservation, now allowing users to align orthologs only or query each species network independently. Bin-free training was applied to 23 primary species, and in addition, networks were

generated for all remaining 618 species in InParanoiDB 9. Accompanying these advancements, FunCoup 6 features a new redesigned website, together with updated API functionalities, and represents a pivotal step forward in functional genomics research, offering unique capabilities for exploring the complex landscape of protein interactions.

- Szklarczyk et al. (2022) — STRING.

  The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Szklarczyk et al. (2022). *Nucleic Acids Research*. doi:10.1093/nar/gkac1000

  > Much of the complexity within cells arises from functional and regulatory interactions among proteins. The core of these interactions is increasingly known, but novel interactions continue to be discovered, and the information remains scattered across different database resources, experimental modalities and levels of mechanistic detail. The STRING database (https://string-db.org/) systematically collects and integrates protein-protein interactions-both physical interactions as well as functional associations. The data originate from a number of sources: automated text mining of the scientific literature, computational interaction predictions from co-expression, conserved genomic context, databases of interaction experiments and known complexes/pathways from curated sources. All of these interactions are critically assessed, scored, and subsequently automatically transferred to less well-studied organisms using hierarchical orthology information. The data can be accessed via the website, but also programmatically and via bulk downloads. The most recent developments in STRING (version 12.0) are: (i) it is now possible to create, browse and analyze a full interaction network for any novel genome of interest, by submitting its complement of encoded proteins, (ii) the co-expression channel now uses variational auto-encoders to predict interactions, and it covers two new sources, single-cell RNA-seq and experimental proteomics data and (iii) the confidence in each experimentally derived interaction is now estimated based on the detection method used, and communicated to the user in the web-interface. Furthermore, STRING continues to enhance its facilities for functional enrichment analysis, which are now fully available also for user-submitted genomes.

B.11.5 Disease networks and systems biology

- Goh et al. (2007), Barabási et al. (2010), Menche et al. (2015) — the human disease network and network medicine.

  The human disease network. Goh et al. (2007). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.0701361104

  > A network of disorders and disease genes linked by known disorder-gene associations offers a platform to explore in a single graph-theoretic framework all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules. We find that essential human genes are likely to encode hub proteins and are expressed widely in most tissues. This suggests that disease genes also would play a central role in the human interactome. In contrast, we find that the vast majority of disease genes are nonessential and show no tendency to encode hub proteins, and their expression pattern indicates that they are localized in the functional periphery of the network. A selection-based model explains the observed difference between essential and disease genes and also suggests that diseases caused by somatic mutations should not be peripheral, a prediction we confirm for cancer genes.

  Network medicine: a network-based approach to human disease. Barabási et al. (2010). *Nature Reviews Genetics*. doi:10.1038/nrg2918

  > Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances in this direction are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified

by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases.

Uncovering disease-disease relationships through the incomplete interactome. Menche et al. (2015). *Science*. doi:10.1126/science.1257601

> According to the disease module hypothesis, the cellular components associated with a disease segregate in the same neighborhood of the human interactome, the map of biologically relevant molecular interactions. Yet, given the incompleteness of the interactome and the limited knowledge of disease-associated genes, it is not obvious if the available data have sufficient coverage to map out modules associated with each disease. Here we derive mathematical conditions for the identifiability of disease modules and show that the network-based location of each disease module determines its pathobiological relationship to other diseases. For example, diseases with overlapping network modules show significant coexpression patterns, symptom similarity, and comorbidity, whereas diseases residing in separated network neighborhoods are phenotypically distinct. These tools represent an interactome-based platform to predict molecular commonalities between phenotypically related diseases, even if they do not share primary disease genes.

- Köhler et al. (2008), Vanunu et al. (2010) — random walks and network propagation.

Walking the Interactome for Prioritization of Candidate Disease Genes. Köhler et al. (2008). *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2008.02.013

> The identification of genes associated with hereditary disorders has contributed to improving medical care and to a better understanding of gene functions, interactions, and pathways. However, there are well over 1500 Mendelian disorders whose molecular basis remains unknown. At present, methods such as linkage analysis can identify the chromosomal region in which unknown disease genes are located, but the regions could contain up to hundreds of candidate genes. In this work, we present a method for prioritization of candidate genes by use of a global network distance measure, random walk analysis, for definition of similarities in protein-protein interaction networks. We tested our method on 110 disease-gene families with a total of 783 genes and achieved an area under the ROC curve of up to 98% on simulated linkage intervals of 100 genes surrounding the disease gene, significantly outperforming previous methods based on local distance measures. Our results not only provide an improved tool for positional-cloning projects but also add weight to the assumption that phenotypically similar diseases are associated with disturbances of subnetworks within the larger protein interactome that extend beyond the disease proteins themselves.

Associating Genes and Protein Complexes with Disease via Network Propagation. Vanunu et al. (2010). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000641

> A fundamental challenge in human health is the identification of disease-causing genes. Recently, several studies have tackled this challenge via a network-based approach, motivated by the observation that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions. However, most of these approaches use only local network information in the inference process and are restricted to inferring single gene associations. Here, we provide a global, network-based method for prioritizing disease genes and inferring protein complex associations, which we call PRINCE. The method is based on formulating constraints on the prioritization function that relate to its smoothness over the network and usage of prior information. We exploit this function to predict not only genes but also protein complex associations with a disease of interest. We test our method on gene-disease association data, evaluating both the prioritization achieved and the protein complexes inferred. We show that our method outperforms extant approaches in both tasks. Using data on 1,369 diseases from the OMIM knowledgebase, our method is able (in a cross validation setting) to rank the true causal gene first for 34% of the diseases, and infer 139 disease-related complexes that are highly coherent in terms of the function, expression and conservation of their member proteins. Importantly, we apply our method to study three multi-factorial diseases for which some causal genes have been found already: prostate cancer, alzheimer and type 2 diabetes mellitus. PRINCE's predictions for these diseases highly match the known literature, suggesting several novel causal genes and protein complexes for further investigation.

- Ideker et al. (2002), Lee et al. (2008) — active subnetworks and phenotype prediction.

Discovering regulatory and signalling circuits in molecular interaction networks. Ideker et al. (2002). *Bioinformatics*. doi:10.1093/bioinformatics/18.suppl_1.s233

> Motivation In model organisms such as yeast, large databases of protein-protein and protein-DNA interactions have become an extremely important resource for the study of protein function, evolution, and gene regulatory dynamics. In this paper we demonstrate that by integrating these interactions with widely-available mRNA expression data, it is possible to generate concrete hypotheses for the underlying mechanisms governing the observed changes in gene expression. To perform this integration systematically and at large scale, we introduce an approach for screening a molecular interaction network to identify active subnetworks, i.e., connected regions of the network that show significant changes in expression over particular subsets of conditions. The method we present here combines a rigorous statistical measure for scoring subnetworks with a search algorithm for identifying subnetworks with high score. Results We evaluated our procedure on a small network of 332 genes and 362 interactions and a large network of 4160 genes containing all 7462 protein-protein and protein-DNA interactions in the yeast public databases. In the case of the small network, we identified five significant subnetworks that covered 41 out of 77 (53%) of all significant changes in expression. Both network analyses returned several top-scoring subnetworks with good correspondence to known regulatory mechanisms in the literature. These results demonstrate how large-scale genomic approaches may be used to uncover signalling and regulatory pathways in a systematic, integrative fashion.

A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Lee et al. (2008). *Nature Genetics*. doi:10.1038/ng.2007.70

> The fundamental aim of genetics is to understand how an organism's phenotype is determined by its genotype, and implicit in this is predicting how changes in DNA sequence alter phenotypes. A single network covering all the genes of an organism might guide such predictions down to the level of individual cells and tissues. To validate this approach, we computationally generated a network covering most C. elegans genes and tested its predictive capacity. Connectivity within this network predicts essentiality, identifying this relationship as an evolutionarily conserved biological principle. Critically, the network makes tissue-specific predictions-we accurately identify genes for most systematically assayed loss-of-function phenotypes, which span diverse cellular and developmental processes. Using the network, we identify 16 genes whose inactivation suppresses defects in the retinoblastoma tumor suppressor pathway, and we successfully predict that the dystrophin complex modulates EGF signaling. We conclude that an analogous network for human genes might be similarly predictive and thus facilitate identification of disease genes and rational therapeutic targets.

- Kitano (2002) — systems biology.

Systems Biology: A Brief Overview. Kitano (2002). *Science*. doi:10.1126/science.1069492

> To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. Properties of systems, such as robustness, emerge as central issues, and understanding these properties may have an impact on the future of medicine. However, many breakthroughs in experimental devices, advanced software, and analytical methods are required before the achievements of systems biology can live up to their much-touted potential.

- Edwards and Palsson (2000), Ibarra et al. (2002) — flux balance analysis of *E. coli*.

The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Edwards & Palsson (2000). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.97.10.5528

> The Escherichia coli MG1655 genome has been completely sequenced. The annotated sequence, biochemical information, and other information were used to reconstruct the E. coli metabolic map. The stoichiometric coefficients for each metabolic enzyme in the E. coli metabolic map were assembled to construct a genome-specific stoichiometric matrix. The E. coli stoichiometric matrix was used to define the system's characteristics and the capabilities of E. coli metabolism. The effects of gene deletions in the central metabolic pathways on the ability of the in silico metabolic network to support growth were assessed, and the in silico predictions were compared with experimental observations. It was shown that based on stoichiometric and capacity constraints the in silico analysis was able to qualitatively predict the growth potential of mutant strains in 86% of the cases examined. Herein, it is demonstrated that the synthesis of in silico metabolic genotypes based on genomic, biochemical, and strain-specific information is possible, and that systems analysis methods are available to analyze and interpret the metabolic phenotype.

Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Ibarra et al. (2002). *Nature*. doi:10.1038/nature01149

> Annotated genome sequences can be used to reconstruct whole-cell metabolic networks. These metabolic networks can be modelled and analysed (computed) to study complex biological functions. In particular, constraints-based in silico models have been used to calculate optimal growth rates on common carbon substrates, and the results were found to be consistent with experimental data under many but not all conditions. Optimal biological functions are acquired through an evolutionary process. Thus, incorrect predictions of in silico models based on optimal performance criteria may be due to incomplete adaptive evolution under the conditions examined. Escherichia coli K-12 MG1655 grows sub-optimally on glycerol as the sole carbon source. Here we show that when placed under growth selection pressure, the growth rate of E. coli on glycerol reproducibly evolved over 40 days, or about 700 generations, from a sub-optimal value to the optimal growth rate predicted from a whole-cell in silico model. These results open the possibility of using adaptive evolution of entire metabolic networks to realize metabolic states that have been determined a priori based on in silico analysis.

- Thiele and Palsson (2010) — building genome-scale metabolic models.

A protocol for generating a high-quality genome-scale metabolic reconstruction. Thiele & Palsson (2010). *Nature Protocols*. doi:10.1038/nprot.2009.203

> Network reconstructions are a common denominator in systems biology. Bottom-up metabolic network reconstructions have been developed over the last 10 years. These reconstructions represent structured knowledge bases that abstract pertinent information on the biochemical transformations taking place within specific target organisms. The conversion of a reconstruction into a mathematical format facilitates a myriad of computational biological studies, including evaluation of network content, hypothesis testing and generation, analysis of phenotypic characteristics and metabolic engineering. To date, genome-scale metabolic reconstructions for more than 30 organisms have been published and this number is expected to increase rapidly. However, these reconstructions differ in quality and coverage that may minimize their predictive potential and use as knowledge bases. Here we present a comprehensive protocol describing each step necessary to build a high-quality genome-scale metabolic reconstruction, as well as the common trials and tribulations. Therefore, this protocol provides a helpful manual for all stages of the reconstruction process.

- Brunk et al. (2018), Robinson et al. (2020) — human metabolic models.

Recon3D enables a three-dimensional view of gene variation in human metabolism. Brunk et al. (2018). *Nature Biotechnology*. doi:10.1038/nbt.4072

> Genome-scale network reconstructions have helped uncover the molecular basis of metabolism. Here we present Recon3D, a computational resource that includes three-dimensional (3D) metabolite and protein structure data and enables integrated analyses of metabolic functions in humans. We use Recon3D to functionally characterize mutations associated with disease, and identify metabolic response signatures that are caused by exposure to certain drugs. Recon3D represents the most comprehensive human metabolic network model to date, accounting for 3,288 open reading frames (representing 17% of functionally annotated human genes), 13,543 metabolic reactions involving 4,140 unique metabolites, and 12,890 protein structures. These data provide a unique resource for investigating molecular mechanisms of human metabolism. Recon3D is available at http://vmh.life.

An atlas of human metabolism. Robinson et al. (2020). *Science Signaling*. doi:10.1126/scisignal.aaz1482

> Genome-scale metabolic models (GEMs) are valuable tools to study metabolism and provide a scaffold for the integrative analysis of omics data. Researchers have developed increasingly comprehensive human GEMs, but the disconnect among different model sources and versions impedes further progress. We therefore integrated and extensively curated the most recent human metabolic models to construct a consensus GEM, Human1. We demonstrated the versatility of Human1 through the generation and analysis of cell- and tissue-specific models using transcriptomic, proteomic, and kinetic data. We also present an accompanying web portal, Metabolic Atlas (https://www.metabolicatlas.org/), which facilitates further exploration and visualization of Human1 content. Human1 was created using a version-controlled, open-source model development framework to enable community-driven curation and refinement. This framework allows Human1 to be an evolving shared resource for future studies of human health and disease.

B.11.6 Prediction

- Chothia and Janin (1975), Jones and Thornton (1996), Bogan and Thorn (1998) — principles of interfaces and hot spots.

Principles of protein–protein recognition. Chothia & Janin (1975). *Nature*. doi:10.1038/256705a0

> The formation of the protein-protein interface by the insulin dimer, the trypsin-PTI complex and the alphabeta oxyhaemoglobin dimer removes 1,130-1,720 A2 of accessible surface from contact with water. The residues forming the interface are close packed: each occupies the same volume as it does in crystals of amino acids. These results indicate that hydrophobicity is the major factor stabilising protein-protein association, while complementarily plays a selective role in deciding which proteins may associate.

Principles of protein-protein interactions.. Jones & Thornton (1996). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.93.1.13

> This review examines protein complexes in the Brookhaven Protein Databank to gain a better understanding of the principles governing the interactions involved in protein-protein recognition. The factors that influence the formation of protein-protein complexes are explored in four different types of protein-protein complexes--homodimeric proteins, heterodimeric proteins, enzyme-inhibitor complexes, and antibody-protein complexes. The comparison between the complexes highlights differences that reflect their biological roles.

Anatomy of hot spots in protein interfaces. Bogan & Thorn (1998). *Journal of Molecular Biology*. doi:10.1006/jmbi.1998.1843

> Binding of one protein to another is involved in nearly all biological functions, yet the principles governing the interaction of proteins are not fully understood. To analyze the contributions of individual amino acid residues in protein-protein binding we have compiled a database of 2325 alanine mutants for which the change in free energy of binding upon mutation to alanine has been measured (available at http://motorhead. ucsf.edu/thorn/hotspot). Our analysis shows that at the level of side-chains there is little correlation between buried surface area and free energy of binding. We find that the free energy of binding is not evenly distributed across interfaces; instead, there are hot spots of binding energy made up of a small subset of residues in the dimer interface. These hot spots are enriched in tryptophan, tyrosine and arginine, and are surrounded by energetically less important residues that most likely serve to occlude bulk solvent from the hot spot. Occlusion of solvent is found to be a necessary condition for highly energetic interactions.

- Burger and Nimwegen (2008), Bitbol et al. (2016) — pairing interaction partners with coevolution.

Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. Burger & van Nimwegen (2008). *Molecular Systems Biology*. doi:10.1038/msb4100203

> Accurate and large-scale prediction of protein-protein interactions directly from amino-acid sequences is one of the great challenges in computational biology. Here we present a new Bayesian network method that predicts interaction partners using only multiple alignments of amino-acid sequences of interacting protein domains, without tunable parameters, and without the need for any training examples. We first apply the method to bacterial two-component systems and comprehensively reconstruct two-component signaling networks across all sequenced bacteria. Comparisons of our predictions with known interactions show that our method infers interaction partners genome-wide with high accuracy. To demonstrate the general applicability of our method we show that it also accurately predicts interaction partners in a recent dataset of polyketide synthases. Analysis of the predicted genome-wide two-component signaling networks shows that cognates (interacting kinase/regulator pairs, which lie adjacent on the genome) and orphans (which lie isolated) form two relatively independent components of the signaling network in each genome. In addition, while most genes are predicted to have only a small number of interaction partners, we find that 10% of orphans form a separate class of 'hub' nodes that distribute and integrate signals to and from up to tens of different interaction partners.

Inferring interaction partners from protein sequences. Bitbol et al. (2016). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1606762113

> Specific protein-protein interactions are crucial in the cell, both to ensure the formation and stability of multiprotein complexes and to enable signal transduction in various pathways. Functional interactions between proteins result in coevolution between the interaction partners, causing their sequences to be correlated. Here we exploit these correlations to accurately identify, from sequence data alone, which proteins are specific interaction partners. Our general approach, which employs a pairwise maximum entropy model to infer couplings between residues, has been successfully used to predict the 3D structures of proteins from sequences. Thus inspired, we introduce an iterative algorithm to predict specific interaction partners from two protein families whose members are known to interact. We first assess the algorithm's performance on histidine kinases and response regulators from bacterial two-component signaling systems. We obtain a striking 0.93 true positive fraction on our complete dataset without any a priori knowledge of interaction partners, and we uncover the origin of this success. We then apply the algorithm to proteins from ATP-binding cassette (ABC) transporter complexes, and obtain accurate predictions in these systems as well. Finally, we present two metrics that accurately distinguish interacting protein families from noninteracting ones, using only sequence data.

- Ovchinnikov et al. (2014), Hopf et al. (2014) — complexes from coevolution.

Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Ovchinnikov et al. (2014). *eLife*. doi:10.7554/elife.02030

> Do the amino acid sequence identities of residues that make contact across protein interfaces covary during evolution? If so, such covariance could be used to predict contacts across interfaces and assemble models of biological complexes. We find that residue pairs identified using a pseudo-likelihood-based method to covary across protein-protein interfaces in the 50S ribosomal unit and 28 additional bacterial protein complexes with known structure are almost always in contact in the complex, provided that the number of aligned sequences is greater than the average length of the two proteins. We use this method to make subunit contact predictions for an additional 36 protein complexes with unknown structures, and present models based on these predictions for the tripartite ATP-independent periplasmic (TRAP) transporter, the tripartite efflux system, the pyruvate formate lyase-activating enzyme complex, and the methionine ABC transporter.DOI: http://dx.doi.org/10.7554/eLife.02030.001.

Sequence co-evolution gives 3D contacts and structures of protein complexes. Hopf et al. (2014). *eLife*. doi:10.7554/elife.03430

> Protein-protein interactions are fundamental to many biological processes. Experimental screens have identified tens of thousands of interactions, and structural biology has provided detailed functional insight for select 3D protein complexes. An alternative rich source of information about protein interactions is the evolutionary sequence record. Building on earlier work, we show that analysis of correlated evolutionary sequence changes across proteins identifies residues that are close in space with sufficient accuracy to determine the three-dimensional structure of the protein complexes. We evaluate prediction performance in blinded tests on 76 complexes of known 3D structure, predict protein-protein contacts in 32 complexes of unknown structure, and demonstrate how evolutionary couplings can be used to distinguish between interacting and non-interacting protein pairs in a large complex. With the current growth of sequences, we expect that the method can be generalized to genome-wide elucidation of protein-protein interaction networks and used for interaction predictions at residue resolution.

- Cong et al. (2019), Green et al. (2021) — proteome-wide coevolution screens.

Protein interaction networks revealed by proteome coevolution. Cong et al. (2019). *Science*. doi:10.1126/science.aaw6718

> Residue-residue coevolution has been observed across a number of protein-protein interfaces, but the extent of residue coevolution between protein families on the whole-proteome scale has not been systematically studied. We investigate coevolution between 5.4 million pairs of proteins in Escherichia coli and between 3.9 millions pairs in Mycobacterium tuberculosis We find strong coevolution for binary complexes involved in metabolism and weaker coevolution for larger complexes playing roles in genetic information processing. We take advantage of this coevolution, in combination with structure modeling, to predict protein-protein interactions (PPIs) with an accuracy that benchmark studies suggest is considerably higher than that of proteome-wide two-hybrid and mass spectrometry screens. We identify hundreds of previously uncharacterized PPIs in E. coli and M. tuberculosis that both add components to known protein complexes and networks and establish the existence of new ones.

Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. Green et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-21636-z

> Increasing numbers of protein interactions have been identified in high-throughput experiments, but only a small proportion have solved structures. Recently, sequence coevolution-based approaches have led to a breakthrough in predicting monomer protein structures and protein interaction interfaces. Here, we address the challenges of large-scale interaction prediction at residue resolution with a fast alignment concatenation method and a probabilistic score for the interaction of residues. Importantly, this method (EVcomplex2) is able to assess the likelihood of a protein interaction, as we show here applied to large-scale experimental datasets where the pairwise interactions are unknown. We predict 504 interactions de novo in the E. coli membrane proteome, including 243 that are newly discovered. While EVcomplex2 does not require available structures, coevolving residue pairs can be used to produce structural models of protein interactions, as done here for membrane complexes including the Flagellar Hook-Filament Junction and the Tol/Pal complex.

- Humphreys et al. (2021), Humphreys et al. (2024) — computed structures of complexes in yeast and pathogens.

Computed structures of core eukaryotic protein complexes. Humphreys et al. (2021). *Science*. doi:10.1126/science.abm4805

> Protein-protein interactions play critical roles in biology, but the structures of many eukaryotic protein complexes are unknown, and there are likely many interactions not yet identified. We take advantage of advances in proteome-wide amino acid coevolution analysis and deep-learning–based structure modeling to systematically identify and build accurate models of core eukaryotic protein complexes within the Saccharomyces cerevisiae proteome. We use a combination of RoseTTAFold and AlphaFold to screen through paired multiple sequence alignments for 8.3 million pairs of yeast proteins, identify 1505 likely to interact, and build structure models for 106 previously unidentified assemblies and 806 that have not been structurally characterized. These complexes, which have as many as five subunits, play roles in almost all key processes in eukaryotic cells and provide broad insights into biological function.

Protein interactions in human pathogens revealed through deep learning. Humphreys et al. (2024). *Nature Microbiology*. doi:10.1038/s41564-024-01791-x

> Identification of bacterial protein-protein interactions and predicting the structures of these complexes could aid in the understanding of pathogenicity mechanisms and developing treatments for infectious diseases. Here we developed RoseTTAFold2-Lite, a rapid deep learning model that leverages residue-residue coevolution and protein structure prediction to systematically identify and structurally characterize protein-protein interactions at the proteome-wide scale. Using this pipeline, we searched through 78 million pairs of proteins across 19 human bacterial pathogens and identified 1,923 confidently predicted complexes involving essential genes and 256 involving virulence factors. Many of these complexes were not previously known; we experimentally tested 12 such predictions, and half of them were validated. The predicted interactions span core metabolic and virulence pathways ranging from post-transcriptional modification to acid neutralization to outer-membrane machinery and should contribute to our understanding of the biology of these important pathogens and the design of drugs to combat them.

- Katchalski-Katzir et al. (1992), Gray et al. (2003), Pierce et al. (2014), Dominguez et al. (2003), Kozakov et al. (2017) — protein–protein docking.

Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.. Katchalski-Katzir et al. (1992). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.89.6.2195

> A geometric recognition algorithm was developed to identify molecular surface complementarity. It is based on a purely geometric approach and takes advantage of techniques applied in the field of pattern recognition. The algorithm involves an automated procedure including (i) a digital representation of the molecules (derived from atomic coordinates) by three-dimensional discrete functions that distinguishes between the surface and the interior; (ii) the calculation, using Fourier transformation, of a correlation function that assesses the degree of molecular surface overlap and penetration upon relative shifts of the molecules in three dimensions; and (iii) a scan of the relative orientations of the molecules in three dimensions. The algorithm provides a list of correlation values indicating the extent of geometric match

between the surfaces of the molecules; each of these values is associated with six numbers describing the relative position (translation and rotation) of the molecules. The procedure is thus equivalent to a six-dimensional search but much faster by design, and the computation time is only moderately dependent on molecular size. The procedure was tested and validated by using five known complexes for which the correct relative position of the molecules in the respective adducts was successfully predicted. The molecular pairs were deoxyhemoglobin and methemoglobin, tRNA synthetase-tyrosinyl adenylate, aspartic proteinase-peptide inhibitor, and trypsin-trypsin inhibitor. A more realistic test was performed with the last two pairs by using the structures of uncomplexed aspartic proteinase and trypsin inhibitor, respectively. The results are indicative of the extent of conformational changes in the molecules tolerated by the algorithm.

Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. Gray et al. (2003). *Journal of Molecular Biology*. doi:10.1016/s0022-2836(03)00670-3

Protein-protein docking algorithms provide a means to elucidate structural details for presently unknown complexes. Here, we present and evaluate a new method to predict protein-protein complexes from the coordinates of the unbound monomer components. The method employs a low-resolution, rigid-body, Monte Carlo search followed by simultaneous optimization of backbone displacement and side-chain conformations using Monte Carlo minimization. Up to 10(5) independent simulations are carried out, and the resulting "decoys" are ranked using an energy function dominated by van der Waals interactions, an implicit solvation model, and an orientation-dependent hydrogen bonding potential. Top-ranking decoys are clustered to select the final predictions. Small-perturbation studies reveal the formation of binding funnels in 42 of 54 cases using coordinates derived from the bound complexes and in 32 of 54 cases using independently determined coordinates of one or both monomers. Experimental binding affinities correlate with the calculated score function and explain the predictive success or failure of many targets. Global searches using one or both unbound components predict at least 25% of the native residue-residue contacts in 28 of the 32 cases where binding funnels exist. The results suggest that the method may soon be useful for generating models of biologically important complexes from the structures of the isolated components, but they also highlight the challenges that must be met to achieve consistent and accurate prediction of protein-protein interactions.

ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. Pierce et al. (2014). *Bioinformatics*. doi:10.1093/bioinformatics/btu097

Protein-protein interactions are essential to cellular and immune function, and in many cases, because of the absence of an experimentally determined structure of the complex, these interactions must be modeled to obtain an understanding of their molecular basis. We present a user-friendly protein docking server, based on the rigid-body docking programs ZDOCK and M-ZDOCK, to predict structures of protein-protein complexes and symmetric multimers. With a goal of providing an accessible and intuitive interface, we provide options for users to guide the scoring and the selection of output models, in addition to dynamic visualization of input structures and output docking models. This server enables the research community to easily and quickly produce structural models of protein-protein complexes and symmetric multimers for their own analysis. Availability The ZDOCK server is freely available to all academic and non-profit users at: http://zdock.umassmed.edu. No registration is required.

HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. Dominguez et al. (2003). *Journal of the American Chemical Society*. doi:10.1021/ja026939x

The structure determination of protein-protein complexes is a rather tedious and lengthy process, by both NMR and X-ray crystallography. Several methods based on docking to study protein complexes have also been well developed over the past few years. Most of these approaches are not driven by experimental data but are based on a combination of energetics and shape complementarity. Here, we present an approach called HADDOCK (High Ambiguity Driven protein-protein Docking) that makes use of biochemical and/or biophysical interaction data such as chemical shift perturbation data resulting from NMR titration experiments or mutagenesis data. This information is introduced as Ambiguous Interaction Restraints (AIRs) to drive the docking process. An AIR is defined as an ambiguous distance between all residues shown to be involved in the interaction. The accuracy of our approach is demonstrated with three molecular complexes. For two of these complexes, for which both the complex and the free protein structures have been solved, NMR titration data were available. Mutagenesis data were used in the last example. In all cases, the best structures generated by HADDOCK, that is, the structures with the lowest

intermolecular energies, were the closest to the published structure of the respective complexes (within 2.0 A backbone RMSD).

The ClusPro web server for protein–protein docking. Kozakov et al. (2017). *Nature Protocols*. doi:10.1038/nprot.2016.169

> The ClusPro server (https://cluspro.org) is a widely used tool for protein-protein docking. The server provides a simple home page for basic use, requiring only two files in Protein Data Bank (PDB) format. However, ClusPro also offers a number of advanced options to modify the search; these include the removal of unstructured protein regions, application of attraction or repulsion, accounting for pairwise distance restraints, construction of homo-multimers, consideration of small-angle X-ray scattering (SAXS) data, and location of heparin-binding sites. Six different energy functions can be used, depending on the type of protein. Docking with each energy parameter set results in ten models defined by centers of highly populated clusters of low-energy docked structures. This protocol describes the use of the various options, the construction of auxiliary restraints files, the selection of the energy parameters, and the analysis of the results. Although the server is heavily used, runs are generally completed in <4 h.

- Janin et al. (2003), Basu and Wallner (2016) — CAPRI and DockQ.

CAPRI: A Critical Assessment of PRedicted Interactions. Janin et al. (2003). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.10381

> CAPRI is a communitywide experiment to assess the capacity of protein-docking methods to predict protein-protein interactions. Nineteen groups participated in rounds 1 and 2 of CAPRI and submitted blind structure predictions for seven protein-protein complexes based on the known structure of the component proteins. The predictions were compared to the unpublished X-ray structures of the complexes. We describe here the motivations for launching CAPRI, the rules that we applied to select targets and run the experiment, and some conclusions that can already be drawn. The results stress the need for new scoring functions and for methods handling the conformation changes that were observed in some of the target systems. CAPRI has already been a powerful drive for the community of computational biologists who development docking algorithms. We hope that this issue of Proteins will also be of interest to the community of structural biologists, which we call upon to provide new targets for future rounds of CAPRI, and to all molecular biologists who view protein-protein recognition as an essential process.

DockQ: A Quality Measure for Protein-Protein Docking Models. Basu & Wallner (2016). *PLOS ONE*. doi:10.1371/journal.pone.0161879

> The state-of-the-art to assess the structural quality of docking models is currently based on three related yet independent quality measures: Fnat, LRMS, and iRMS as proposed and standardized by CAPRI. These quality measures quantify different aspects of the quality of a particular docking model and need to be viewed together to reveal the true quality, e.g. a model with relatively poor LRMS (>10Å) might still qualify as 'acceptable' with a descent Fnat (>0.50) and iRMS (<3.0Å). This is also the reason why the so called CAPRI criteria for assessing the quality of docking models is defined by applying various ad-hoc cutoffs on these measures to classify a docking model into the four classes: Incorrect, Acceptable, Medium, or High quality. This classification has been useful in CAPRI, but since models are grouped in only four bins it is also rather limiting, making it difficult to rank models, correlate with scoring functions or use it as target function in machine learning algorithms. Here, we present DockQ, a continuous protein-protein docking model quality measure derived by combining Fnat, LRMS, and iRMS to a single score in the range [0, 1] that can be used to assess the quality of protein docking models. By using DockQ on CAPRI models it is possible to almost completely reproduce the original CAPRI classification into Incorrect, Acceptable, Medium and High quality. An average PPV of 94% at 90% Recall demonstrating that there is no need to apply predefined ad-hoc cutoffs to classify docking models. Since DockQ recapitulates the CAPRI classification almost perfectly, it can be viewed as a higher resolution version of the CAPRI classification, making it possible to estimate model quality in a more quantitative way using Z-scores or sum of top ranked models, which has been so valuable for the CASP community. The possibility to directly correlate a quality measure to a scoring function has been crucial for the development of scoring functions for protein structure prediction, and DockQ should be useful in a similar development in the protein docking field. DockQ is available at http://github.com/bjornwallner/DockQ/.

- Evans et al. (2021) — AlphaFold-Multimer.

Protein complex prediction with AlphaFold-Multimer. Evans et al. (2021). *bioRxiv (preprint)*. doi:10.1101/2021.10.04.463034

> While the vast majority of well-structured single protein chains can now be predicted to high accuracy due to the recent AlphaFold [1] model, the prediction of multi-chain protein complexes remains a challenge in many cases. In this work, we demonstrate that an AlphaFold model trained specifically for multimeric inputs of known stoichiometry, which we call AlphaFold-Multimer, significantly increases accuracy of predicted multimeric interfaces over input-adapted single-chain AlphaFold while maintaining high intra-chain accuracy. On a benchmark dataset of 17 heterodimer proteins without templates (introduced in [2]) we achieve at least medium accuracy (DockQ [3] ≥ 0.49) on 13 targets and high accuracy (DockQ ≥ 0.8) on 7 targets, compared to 9 targets of at least medium accuracy and 4 of high accuracy for the previous state of the art system (an AlphaFold-based system from [2]). We also predict structures for a large dataset of 4,446 recent protein complexes, from which we score all non-redundant interfaces with low template identity. For heteromeric interfaces we successfully predict the interface (DockQ ≥ 0.23) in 70% of cases, and produce high accuracy predictions (DockQ ≥ 0.8) in 26% of cases, an improvement of +27 and +14 percentage points over the flexible linker modification of AlphaFold [4] respectively. For homomeric inter-faces we successfully predict the interface in 72% of cases, and produce high accuracy predictions in 36% of cases, an improvement of +8 and +7 percentage points respectively.

- Pozzati et al. (2021), Bryant, Pozzati, Zhu, et al. (2022) — folding and docking, and large complexes.

Limits and potential of combined folding and docking. Pozzati et al. (2021). *Bioinformatics*. doi:10.1093/bioinformatics/btab760

> Motivation In the last decade, de novo protein structure prediction accuracy for individual proteins has improved significantly by utilising deep learning (DL) methods for harvesting the co-evolution information from large multiple sequence alignments (MSAs). The same approach can, in principle, also be used to extract information about evolutionary-based contacts across protein-protein interfaces. However, most earlier studies have not used the latest DL methods for inter-chain contact distance prediction. This article introduces a fold-and-dock method based on predicted residue-residue distances with trRosetta. Results The method can simultaneously predict the tertiary and quaternary structure of a protein pair, even when the structures of the monomers are not known. The straightforward application of this method to a standard dataset for protein-protein docking yielded limited success. However, using alternative methods for generating MSAs allowed us to dock accurately significantly more proteins. We also introduced a novel scoring function, PconsDock, that accurately separates 98% of correctly and incorrectly folded and docked proteins. The average performance of the method is comparable to the use of traditional, template-based or ab initio shape-complementarity-only docking methods. Moreover, the results of conventional and fold-and-dock approaches are complementary, and thus a combined docking pipeline could increase overall docking success significantly. This methodology contributed to the best model for one of the CASP14 oligomeric targets, H1065. Availability and implementation All scripts for predictions and analysis are available from https://github.com/ElofssonLab/bioinfo-toolbox/ and https://gitlab.com/ElofssonLab/benchmark5/. All models joined alignments, and evaluation results are available from the following figshare repository https://doi.org/10.6084/m9.figshare.14654886.v2. Supplementary information Supplementary data are available at Bioinformatics online.

Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Bryant et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-33729-4

> AlphaFold can predict the structure of single- and multiple-chain proteins with very high accuracy. However, the accuracy decreases with the number of chains, and the available GPU memory limits the size of protein complexes which can be predicted. Here we show that one can predict the structure of large complexes starting from predictions of subcomponents. We assemble 91 out of 175 complexes with 10-30 chains from predicted subcomponents using Monte Carlo tree search, with a median TM-score of 0.51. There are 30 highly accurate complexes (TM-score ≥0.8, 33% of complete assemblies). We create a scoring function, mpDockQ, that can distinguish if assemblies are complete and predict their accuracy. We find that complexes containing symmetry are accurately assembled, while asymmetrical complexes remain challenging. The method is freely available and accesible as a Colab notebook https://colab.research.google.com/github/patrickbryant1/MoLPC/blob/master/MoLPC.ipynb .

- Zhu et al. (2023) — evaluation of AlphaFold-Multimer.

Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. Zhu et al. (2023). *Bioinformatics*. doi:10.1093/bioinformatics/btad424

Motivation Despite near-experimental accuracy on single-chain predictions, there is still scope for improvement among multimeric predictions. Methods like AlphaFold-Multimer and FoldDock can accurately model dimers. However, how well these methods fare on larger complexes is still unclear. Further, evaluation methods of the quality of multimeric complexes are not well established. Results We analysed the performance of AlphaFold-Multimer on a homology-reduced dataset of homo- and heteromeric protein complexes. We highlight the differences between the pairwise and multi-interface evaluation of chains within a multimer. We describe why certain complexes perform well on one metric (e.g. TM-score) but poorly on another (e.g. DockQ). We propose a new score, Predicted DockQ version 2 (pDockQ2), to estimate the quality of each interface in a multimer. Finally, we modelled protein complexes (from CORUM) and identified two highly confident structures that do not have sequence homology to any existing structures. Availability and implementation All scripts, models, and data used to perform the analysis in this study are freely available at https://gitlab.com/ElofssonLab/afm-benchmark.

B.12 Machine learning

Further papers for Chapter 12.

⋆ the 2021 version of the exam was based on Sapoval et al. (2022) together with Pevsner (2015). In the HTML version, click *Abstract* to show the abstract.

B.12.1 Reviews

- Tarca et al. (2007) — machine learning and its applications to biology.

  Machine Learning and Its Applications to Biology. Tarca et al. (2007). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.0030116

  The term machine learning refers to a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data. Two facets of mechanization should be acknowledged when considering machine learning in broad terms. Firstly, it is intended that the classification and prediction tasks can be accomplished by a suitably programmed computing machine. That is, the product of machine learning is a classifier that can be feasibly used on available hardware. Secondly, it is intended that the creation of the classifier should itself be highly mechanized, and should not involve too much human input. This second facet is inevitably vague, but the basic objective is that the use of automatic algorithm construction methods can minimize the possibility that human biases could affect the selection and performance of the algorithm. Both the creation of the algorithm and its operation to classify objects or predict events are to be based on concrete, observable data. The history of relations between biology and the field of machine learning is long and complex. An early technique [1] for machine learning called the perceptron constituted an attempt to model actual neuronal behavior, and the field of artificial neural network (ANN) design emerged from this attempt. Early work on the analysis of translation initiation sequences [2] employed the perceptron to define criteria for start sites in Escherichia coli. Further artificial neural network architectures such as the adaptive resonance theory (ART) [3] and neocognitron [4] were inspired from the organization of the visual nervous system. In the intervening years, the flexibility of machine learning techniques has grown along with mathematical frameworks for measuring their reliability, and it is natural to hope that machine learning methods will improve the efficiency of discovery and understanding in the mounting volume and complexity of biological data. This tutorial is structured in four main components. Firstly, a brief section reviews definitions and mathematical prerequisites. Secondly, the field of supervised learning is described. Thirdly, methods of unsupervised learning are reviewed. Finally, a section reviews methods and examples as implemented in the open source data analysis and visualization language R (http://www.r-project.org).

- Ching et al. (2018) — opportunities and obstacles for deep learning in biology and medicine.

  Opportunities and obstacles for deep learning in biology and medicine. Ching et al. (2018). *Journal of The Royal Society Interface*. doi:10.1098/rsif.2017.0387

> Deep learning describes a class of machine learning algorithms that are capable of combining raw inputs into layers of intermediate features. These algorithms have recently shown impressive results across a variety of domains. Biology and medicine are data-rich disciplines, but the data are complex and often ill-understood. Hence, deep learning techniques may be particularly well suited to solve problems of these fields. We examine applications of deep learning to a variety of biomedical problems-patient classification, fundamental biological processes and treatment of patients-and discuss whether deep learning will be able to transform these tasks or if the biomedical sphere poses unique challenges. Following from an extensive literature review, we find that deep learning has yet to revolutionize biomedicine or definitively resolve any of the most pressing challenges in the field, but promising advances have been made on the prior state of the art. Even though improvements over previous baselines have been modest in general, the recent progress indicates that deep learning methods will provide valuable means for speeding up or aiding human investigation. Though progress has been made linking a specific neural network's prediction to input features, understanding how users should interpret these models to make testable hypotheses about the system under study remains an open challenge. Furthermore, the limited amount of labelled data for training presents problems in some domains, as do legal and privacy constraints on work with sensitive health records. Nonetheless, we foresee deep learning enabling changes at both bench and bedside with the potential to transform several areas of biology and medicine.

- Eraslan et al. (2019) — deep learning for genomics.

  Deep learning: new computational modelling techniques for genomics. Eraslan et al. (2019). *Nature Reviews Genetics*. doi:10.1038/s41576-019-0122-6

  > As a data-driven science, genomics largely utilizes machine learning to capture dependencies in data and derive novel biological hypotheses. However, the ability to extract new insights from the exponentially increasing volume of genomics data requires more expressive machine learning models. By effectively leveraging large data sets, deep learning has transformed fields such as computer vision and natural language processing. Now, it is becoming the method of choice for many genomics modelling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.

B.12.2 Short primers

- Eddy (2004a) — what is a hidden Markov model?

  What is a hidden Markov model?. Eddy (2004). *Nature Biotechnology*. doi:10.1038/nbt1004-1315

- Noble (2006) — what is a support vector machine?

  What is a support vector machine?. Noble (2006). *Nature Biotechnology*. doi:10.1038/nbt1206-1565

  > Support vector machines (SVMs) are becoming popular in a wide variety of biological applications. But, what exactly are SVMs and how do they work? And what are their most promising applications in the life sciences?

- Lever et al. (2016), Lever et al. (2017) — model selection and overfitting; principal component analysis.

  Model selection and overfitting. Lever et al. (2016). *Nature Methods*. doi:10.1038/nmeth.3968

  Principal component analysis. Lever et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4346

B.12.3 Basics and evaluation

- Hobohm et al. (1992), Li and Godzik (2006) — non-redundant data sets.

  Selection of representative protein data sets. Hobohm et al. (1992). *Protein Science*. doi:10.1002/pro.5560010313

  > The Protein Data Bank currently contains about 600 data sets of three-dimensional protein coordinates determined by X-ray crystallography or NMR. There is considerable redundancy in the data base, as many protein pairs are identical or very similar in sequence. However, statistical analyses of protein sequence-structure relations require nonredundant data. We have developed two algorithms to extract from the data base representative sets of protein chains with maximum coverage and minimum redundancy. The first algorithm focuses on optimizing a particular property of the selected proteins and works by successive selection of proteins from an ordered list and exclusion of all neighbors of each selected protein. The other algorithm aims at maximizing the size of the selected set and works by successive thinning out of clusters of similar proteins. Both algorithms are generally applicable to other data bases in which criteria of similarity can be defined and relate to problems in graph theory. The largest nonredundant set extracted from the current release of the Protein Data Bank has 155 protein chains. In this set, no two proteins have sequence similarity higher than a certain cutoff (30% identical residues for aligned subsequences longer than 80 residues), yet all structurally unique protein families are represented. Periodically updated lists of representative data sets are available by electronic mail from the file server "netserv@embl-heidelberg.de." The selection may be useful in statistical approaches to protein folding as well as in the analysis and documentation of the known spectrum of three-dimensional protein structures.

  Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Li & Godzik (2006). *Bioinformatics*. doi:10.1093/bioinformatics/btl158

  > Motivation In 2001 and 2002, we published two papers (Bioinformatics, 17, 282-283, Bioinformatics, 18, 77-82) describing an ultrafast protein sequence clustering program called cd-hit. This program can efficiently cluster a huge protein database with millions of sequences. However, the applications of the underlying algorithm are not limited to only protein sequences clustering, here we present several new programs using the same algorithm including cd-hit-2d, cd-hit-est and cd-hit-est-2d. Cd-hit-2d compares two protein datasets and reports similar matches between them; cd-hit-est clusters a DNA/RNA sequence database and cd-hit-est-2d compares two nucleotide datasets. All these programs can handle huge datasets with millions of sequences and can be hundreds of times faster than methods based on the popular sequence comparison and database search tools, such as BLAST.

- Saito and Rehmsmeier (2015) — precision–recall vs. ROC curves.

  The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Saito & Rehmsmeier (2015). *PLOS ONE*. doi:10.1371/journal.pone.0118432

  > Binary classifiers are routinely evaluated with performance measures such as sensitivity and specificity, and performance is frequently illustrated with Receiver Operating Characteristics (ROC) plots. Alternative measures such as positive predictive value (PPV) and the associated Precision/Recall (PRC) plots are used less frequently. Many bioinformatics studies develop and evaluate classifiers that are to be applied to strongly imbalanced datasets in which the number of negatives outweighs the number of positives significantly. While ROC plots are visually appealing and provide an overview of a classifier's performance across a wide range of specificities, one can ask whether ROC plots could be misleading when applied in imbalanced classification scenarios. We show here that the visual interpretability of ROC plots in the context of imbalanced datasets can be deceptive with respect to conclusions about the reliability of classification performance, owing to an intuitive but wrong interpretation of specificity. PRC plots, on the other hand, can provide the viewer with an accurate prediction of future classification performance due to the fact that they evaluate the fraction of true positives among positive predictions. Our findings have potential implications for the interpretation of a large number of studies that use ROC plots on imbalanced datasets.

- Jones (2019) — setting the standards for machine learning in biology.

  Setting the standards for machine learning in biology. Jones (2019). *Nature Reviews Molecular Cell Biology*. doi:10.1038/s41580-019-0176-5

B.12.4 History and classical methods

- Rosenblatt (1958), Minsky and Papert (1969) — the perceptron and its limits.

  The perceptron: A probabilistic model for information storage and organization in the brain.. Rosenblatt (1958). *Psychological Review*. doi:10.1037/h0042519

  > If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions: 1. How is information about the physical world sensed, or detected, by the biological system? 2. In what form is information stored, or remembered? 3. How does information contained in storage, or in memory, influence recognition and behavior? The first of these questions is in the

  Perceptrons: An Introduction to Computational Geometry. Minsky & Papert (1969). *MIT Press*.

- Rumelhart et al. (1986) — backpropagation.

  Learning representations by back-propagating errors. Rumelhart et al. (1986). *Nature*. doi:10.1038/323533a0

- Stormo et al. (1982) — the perceptron applied to translation initiation sites.

  Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Stormo et al. (1982). *Nucleic Acids Research*. doi:10.1093/nar/10.9.2997

  > We have used a "Perceptron" algorithm to find a weighting function which distinguishes E. coli translational initiation sites from all other sites in a library of over 78,000 nucleotides of mRNA sequence. The "Perceptron" examined sequences as linear representations. The "Perceptron" is more successful at finding gene beginnings than our previous searches using "rules" (see previous paper). We note that the weighting function can find translational initiation sites within sequences that were not included in the training set.

- Qian and Sejnowski (1988), Holley and Karplus (1989), Rost and Sander (1993) — neural networks for secondary structure.

  Predicting the secondary structure of globular proteins using neural network models. Qian & Sejnowski (1988). *Journal of Molecular Biology*. doi:10.1016/0022-2836(88)90564-5

  > We present a new method for predicting the secondary structure of globular proteins based on non-linear neural network models. Network models learn from existing protein structures how to predict the secondary structure of local sequences of amino acids. The average success rate of our method on a testing set of proteins non-homologous with the corresponding training set was 64.3% on three types of secondary structure (alpha-helix, beta-sheet, and coil), with correlation coefficients of C alpha = 0.41, C beta = 0.31 and Ccoil = 0.41. These quality indices are all higher than those of previous methods. The prediction accuracy for the first 25 residues of the N-terminal sequence was significantly better. We conclude from computational experiments on real and artificial structures that no method based solely on local information in the protein sequence is likely to produce significantly better results for non-homologous proteins. The performance of our method of homologous proteins is much better than for non-homologous proteins, but is not as good as simply assuming that homologous sequences have identical structures.

  Protein secondary structure prediction with a neural network.. Holley & Karplus (1989). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.86.1.152

  > A method is presented for protein secondary structure prediction based on a neural network. A training phase was used to teach the network to recognize the relation between secondary structure and amino acid sequences on a sample set of 48 proteins of known structure. On a separate test set of 14 proteins of known structure, the method achieved a maximum overall predictive accuracy of 63% for three states: helix, sheet, and coil. A numerical measure of helix and sheet tendency for each residue was obtained from the calculations. When predictions were filtered to include only the strongest 31% of predictions, the predictive accuracy rose to 79%.

Prediction of Protein Secondary Structure at Better than 70% Accuracy. Rost & Sander (1993). *Journal of Molecular Biology*. doi:10.1006/jmbi.1993.1413

> We have trained a two-layered feed-forward neural network on a non-redundant data base of 130 protein chains to predict the secondary structure of water-soluble proteins. A new key aspect is the use of evolutionary information in the form of multiple sequence alignments that are used as input in place of single sequences. The inclusion of protein family information in this form increases the prediction accuracy by six to eight percentage points. A combination of three levels of networks results in an overall three-state accuracy of 70.8% for globular proteins (sustained performance). If four membrane protein chains are included in the evaluation, the overall accuracy drops to 70.2%. The prediction is well balanced between alpha-helix, beta-strand and loop: 65% of the observed strand residues are predicted correctly. The accuracy in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy. The performance accuracy is verified by a sevenfold cross-validation test, and an additional test on 26 recently solved proteins. Of particular practical importance is the definition of a position-specific reliability index. For half of the residues predicted with a high level of reliability the overall accuracy increases to better than 82%. A further strength of the method is the more realistic prediction of segment length. The protein family prediction method is available for testing by academic researchers via an electronic mail server.

- Rabiner (1989), Viterbi (1967), Baum et al. (1970) — HMMs and their algorithms.

A tutorial on hidden Markov models and selected applications in speech recognition. Rabiner (1989). *Proceedings of the IEEE*. doi:10.1109/5.18626

> This tutorial provides an overview of the basic theory of hidden Markov models (HMMs) as originated by L.E. Baum and T. Petrie (1966) and gives practical details on methods of implementation of the theory along with a description of selected applications of the theory to distinct problems in speech recognition. Results from a number of original sources are combined to provide a single source of acquiring the background required to pursue further this area of research. The author first reviews the theory of discrete Markov chains and shows how the concept of hidden states, where the observation is a probabilistic function of the state, can be used effectively. The theory is illustrated with two simple examples, namely coin-tossing, and the classic balls-in-urns system. Three fundamental problems of HMMs are noted and several practical techniques for solving these problems are given. The various types of HMMs that have been studied, including ergodic as well as left-right models, are described.>

Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Viterbi (1967). *IEEE Transactions on Information Theory*. doi:10.1109/tit.1967.1054010

> The probability of error in decoding an optimal convolutional code transmitted over a memoryless channel is bounded from above and below as a function of the constraint length of the code. For all but pathological channels the bounds are asymptotically (exponentially) tight for rates aboveR_{0}, the computational cutoff rate of sequential decoding. As a function of constraint length the performance of optimal convolutional codes is shown to be superior to that of block codes of the same length, the relative improvement increasing with rate. The upper bound is obtained for a specific probabilistic nonsequential decoding algorithm which is shown to be asymptotically optimum for rates aboveR_{0}and whose performance bears certain similarities to that of sequential decoding algorithms.

A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Baum et al. (1970). *The Annals of Mathematical Statistics*. doi:10.1214/aoms/1177697196

- Churchill (1989), Krogh et al. (1994) — HMMs in biology.

Stochastic models for heterogeneous DNA sequences. Churchill (1989). *Bulletin of Mathematical Biology*. doi:10.1007/bf02458837

> The composition of naturally occurring DNA sequences is often strikingly heterogeneous. In this paper, the DNA sequence is viewed as a stochastic process with local compositional properties determined by the states of a hidden Markov chain. The model used is a discrete-state, discrete-outcome version of a general model for non-stationary time series proposed by Kitagawa (1987). A smoothing algorithm is described which can be used to reconstruct the hidden process and produce graphic displays of the compositional structure of a sequence. The problem of parameter estimation is approached using likelihood methods and an EM algorithm for approximating the maximum likelihood estimate is derived. The methods are

applied to sequences from yeast mitochondrial DNA, human and mouse mitochondrial DNAs, a human X chromosomal fragment and the complete genome of bacteriophage lambda.

Hidden Markov Models in Computational Biology. Krogh et al. (1994). *Journal of Molecular Biology*. doi:10.1006/jmbi.1994.1104

> Hidden Markov Models (HMMs) are applied to the problems of statistical modeling, database searching and multiple sequence alignment of protein families and protein domains. These methods are demonstrated on the globin family, the protein kinase catalytic domain, and the EF-hand calcium binding motif. In each case the parameters of an HMM are estimated from a training set of unaligned sequences. After the HMM is built, it is used to obtain a multiple alignment of all the training sequences. It is also used to search the SWISS-PROT 22 database for other sequences that are members of the given protein family, or contain the given domain. The HMM produces multiple alignments of good quality that agree closely with the alignments produced by programs that incorporate three-dimensional structural information. When employed in discrimination tests (by examining how closely the sequences in a database fit the globin, kinase and EF-hand HMMs), the HMM is able to distinguish members of these families from non-members with a high degree of accuracy. Both the HMM and PROFILESEARCH (a technique used to search for relationships between a protein sequence and multiply aligned sequences) perform better in these tests than PROSITE (a dictionary of sites and patterns in proteins). The HMM appears to have a slight advantage over PROFILESEARCH in terms of lower rates of false negatives and false positives, even though the HMM is trained using only unaligned sequences, whereas PROFILESEARCH requires aligned training sequences. Our results suggest the presence of an EF-hand calcium binding motif in a highly conserved and evolutionary preserved putative intracellular region of 155 residues in the alpha-1 subunit of L-type calcium channels which play an important role in excitation-contraction coupling. This region has been suggested to contain the functional domains that are typical or essential for all L-type calcium channels regardless of whether they couple to ryanodine receptors, conduct ions or both.

- Durbin et al. (1998) — *Biological Sequence Analysis.*

  Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Durbin et al. (1998). *Cambridge University Press*. doi:10.1017/cbo9780511790492

  > Probabilistic models are becoming increasingly important in analysing the huge amount of data being produced by large-scale DNA-sequencing efforts such as the Human Genome Project. For example, hidden Markov models are used for analysing biological sequences, linguistic-grammar-based probabilistic models for identifying RNA secondary structure, and probabilistic evolutionary models for inferring phylogenies of sequences from different organisms. This book gives a unified, up-to-date and self-contained account, with a Bayesian slant, of such methods, and more generally to probabilistic methods of sequence analysis. Written by an interdisciplinary team of authors, it aims to be accessible to molecular biologists, computer scientists, and mathematicians with no formal knowledge of the other fields, and at the same time present the state-of-the-art in this new and highly important field.

- Cortes and Vapnik (1995) — support vector machines.

  Support-vector networks. Cortes & Vapnik (1995). *Machine Learning*. doi:10.1007/bf00994018

- Breiman (2001), Friedman (2001), Chen and Guestrin (2016) — random forests and gradient boosting.

  Random Forests. Breiman (2001). *Machine Learning*. doi:10.1023/a:1010933404324

  Greedy function approximation: A gradient boosting machine.. Friedman (2001). *The Annals of Statistics*. doi:10.1214/aos/1013203451

  > Function estimation/approximation is viewed from the perspective of numerical optimization in function space, rather than parameter space. A connection is made between stagewise additive expansions and steepest-descent minimization. A general gradient descent "boosting" paradigm is developed for additive expansions based on any fitting criterion.Specific algorithms are presented for least-squares, least absolute deviation, and Huber-M loss functions for regression, and multiclass logistic likelihood for classification. Special enhancements are derived for the particular case where the individual additive components are regression trees, and tools for interpreting such "TreeBoost" models are presented. Gradient boosting of regression trees produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than clean data. Connections between this approach and the boosting methods of Freund and Shapire and Friedman, Hastie and Tibshirani are discussed.

XGBoost: A Scalable Tree Boosting System. Chen & Guestrin (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2939672.2939785

> Over 78 million people will suffer from dementia by 2030, emphasizing the need for early identification of patients with mild cognitive impairment (MCI) at risk, and personalized clinical evaluation steps to diagnose potentially reversible causes. Here, we leverage real-world electronic health records in the observational medical outcomes partnership (OMOP) data model to develop machine learning models to predict MCI up to a year in advance of recorded diagnosis. Our experimental results with logistic regression, random forest, and xgboost models trained and evaluated on more than 531K patient visits show random forest model can predict MCI onset with ROC-AUC of 68.2±0.7. We identify the clinical factors mentioned in clinician notes that are most predictive of MCI. Using similar association mining techniques, we develop a data-driven list of clinical procedures commonly ordered in the workup of MCI cases, that could be used as a basis for guidelines and clinical order set templates.

- Hastie et al. (2009) — *The Elements of Statistical Learning* (free PDF).

  The Elements of Statistical Learning. Hastie et al. (2009). *Springer*. https://hastie.su.domains/ElemStatLearn/

B.12.5 Deep learning

- Lecun et al. (1998), Krizhevsky et al. (2017), He et al. (2016) — CNNs, AlexNet and residual networks.

  Gradient-based learning applied to document recognition. Lecun et al. (1998). *Proceedings of the IEEE*. doi:10.1109/5.726791

  > Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional neural networks, which are specifically designed to deal with the variability of 2D shapes, are shown to outperform all other techniques. Real-life document recognition systems are composed of multiple modules including field extraction, segmentation recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure. Two systems for online handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks. A graph transformer network for reading a bank cheque is also described. It uses convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal cheques. It is deployed commercially and reads several million cheques per day.

  ImageNet classification with deep convolutional neural networks. Krizhevsky et al. (2017). *Communications of the ACM*. doi:10.1145/3065386

  > We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully connected layers we employed a recently developed regularization method called ”dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

  Deep Residual Learning for Image Recognition. He et al. (2016). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.90

> Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers - 8× deeper than VGG nets [40] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions1, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

- Hochreiter and Schmidhuber (1997) — LSTM.

  Long Short-Term Memory. Hochreiter & Schmidhuber (1997). *Neural Computation*. doi:10.1162/neco.1997.9.8.1735

  > Learning to store information over extended time intervals by recurrent backpropagation takes a very long time, mostly because of insufficient, decaying error backflow. We briefly review Hochreiter's (1991) analysis of this problem, then address it by introducing a novel, efficient, gradient-based method called long short-term memory (LSTM). Truncating the gradient where this does not do harm, LSTM can learn to bridge minimal time lags in excess of 1000 discrete-time steps by enforcing constant error flow through constant error carousels within special units. Multiplicative gate units learn to open and close access to the constant error flow. LSTM is local in space and time; its computational complexity per time step and weight is O(1). Our experiments with artificial data involve local, distributed, real-valued, and noisy pattern representations. In comparisons with real-time recurrent learning, back propagation through time, recurrent cascade correlation, Elman nets, and neural sequence chunking, LSTM leads to many more successful runs, and learns much faster. LSTM also solves complex, artificial long-time-lag tasks that have never been solved by previous recurrent network algorithms.

- Srivastava et al. (2014) — dropout.

  Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Srivastava et al. (2014). *Journal of Machine Learning Research*. https://jmlr.org/papers/v15/srivastava14a.html

- Vaswani et al. (2017) — the transformer.

  Attention is all you need. Vaswani et al. (2017). *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. https://arxiv.org/abs/1706.03762

- Alipanahi et al. (2015), Zhou and Troyanskaya (2015), Kelley et al. (2016) — CNNs for DNA and RNA sequences.

  Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Alipanahi et al. (2015). *Nature Biotechnology*. doi:10.1038/nbt.3300

  > Knowing the sequence specificities of DNA- and RNA-binding proteins is essential for developing models of the regulatory processes in biological systems and for identifying causal disease variants. Here we show that sequence specificities can be ascertained from experimental data with 'deep learning' techniques, which offer a scalable, flexible and unified computational approach for pattern discovery. Using a diverse array of experimental data and evaluation metrics, we find that deep learning outperforms other state-of-the-art methods, even when training on in vitro data and testing on in vivo data. We call this approach DeepBind and have built a stand-alone software tool that is fully automatic and handles millions of sequences per experiment. Specificities determined by DeepBind are readily visualized as a weighted ensemble of position weight matrices or as a 'mutation map' that indicates how variations affect binding within a specific sequence.

  Predicting effects of noncoding variants with deep learning–based sequence model. Zhou & Troyanskaya (2015). *Nature Methods*. doi:10.1038/nmeth.3547

Identifying functional effects of noncoding variants is a major challenge in human genetics. To predict the noncoding-variant effects de novo from sequence, we developed a deep learning-based algorithmic framework, DeepSEA (http://deepsea.princeton.edu/), that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity. We further used this capability to improve prioritization of functional variants including expression quantitative trait loci (eQTLs) and disease-associated variants.

Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Kelley et al. (2016). *Genome Research*. doi:10.1101/gr.200535.115

The complex language of eukaryotic gene expression remains incompletely understood. Despite the importance suggested by many noncoding variants statistically associated with human disease, nearly all such variants have unknown mechanisms. Here, we address this challenge using an approach based on a recent machine learning advance-deep convolutional neural networks (CNNs). We introduce the open source package Basset to apply CNNs to learn the functional activity of DNA sequences from genomics data. We trained Basset on a compendium of accessible genomic sites mapped in 164 cell types by DNase-seq, and demonstrate greater predictive accuracy than previous methods. Basset predictions for the change in accessibility between variant alleles were far greater for Genome-wide association study (GWAS) SNPs that are likely to be causal relative to nearby SNPs in linkage disequilibrium with them. With Basset, a researcher can perform a single sequencing assay in their cell type of interest and simultaneously learn that cell's chromatin accessibility code and annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility. Thus, Basset offers a powerful computational approach to annotate and interpret the noncoding genome.

- Rives et al. (2021), Elnaggar et al. (2022) — protein language models.

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Rives et al. (2021). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.2016239118

In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In the life sciences, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Protein language modeling at the scale of evolution is a logical step toward predictive and generative artificial intelligence for biology. To this end, we use unsupervised learning to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity. The resulting model contains information about biological properties in its representations. The representations are learned from sequence data alone. The learned representation space has a multiscale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins. Information about secondary and tertiary structure is encoded in the representations and can be identified by linear projections. Representation learning produces features that generalize across a range of applications, enabling state-of-the-art supervised prediction of mutational effect and secondary structure and improving state-of-the-art features for long-range contact prediction.

ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. Elnaggar et al. (2022). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/tpami.2021.3095381

Computational biology and bioinformatics provide vast data gold-mines from protein sequences, ideal for Language Models (LMs) taken from Natural Language Processing (NLP). These LMs reach for new prediction frontiers at low inference costs. Here, we trained two auto-regressive models (Transformer-XL, XLNet) and four auto-encoder models (BERT, Albert, Electra, T5) on data from UniRef and BFD containing up to 393 billion amino acids. The protein LMs (pLMs) were trained on the Summit supercomputer using 5616 GPUs and TPU Pod up-to 1024 cores. Dimensionality reduction revealed that the raw pLM-embeddings from unlabeled data captured some biophysical features of protein sequences. We validated the advantage of using the embeddings as exclusive input for several subsequent tasks: (1) a per-residue (per-token) prediction of protein secondary structure (3-state accuracy Q3=81%-87%); (2) per-protein (pooling) predictions of protein sub-cellular location (ten-state accuracy: Q10=81%) and membrane versus water-soluble (2-state accuracy Q2=91%). For secondary structure, the most informative embeddings (ProtT5) for the first time outperformed the state-of-the-art without multiple sequence alignments (MSAs) or evolutionary information thereby bypassing expensive database searches. Taken together, the results

implied that pLMs learned some of the grammar of the language of life. All our models are available through https://github.com/agemagician/ProtTrans.

- Ji et al. (2021), Dalla-Torre et al. (2024) — DNA language models.

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Ji et al. (2021). *Bioinformatics*. doi:10.1093/bioinformatics/btab083

> Motivation Deciphering the language of non-coding DNA is one of the fundamental problems in genome research. Gene regulatory code is highly complex due to the existence of polysemy and distant semantic relationship, which previous informatics methods often fail to capture especially in data-scarce scenarios. Results To address this challenge, we developed a novel pre-trained bidirectional encoder representation, named DNABERT, to capture global and transferrable understanding of genomic DNA sequences based on up and downstream nucleotide contexts. We compared DNABERT to the most widely used programs for genome-wide regulatory elements prediction and demonstrate its ease of use, accuracy and efficiency. We show that the single pre-trained transformers model can simultaneously achieve state-of-the-art performance on prediction of promoters, splice sites and transcription factor binding sites, after easy fine-tuning using small task-specific labeled data. Further, DNABERT enables direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence motifs and functional genetic variant candidates. Finally, we demonstrate that pre-trained DNABERT with human genome can even be readily applied to other organisms with exceptional performance. We anticipate that the pre-trained DNABERT model can be fined tuned to many other sequence analyses tasks. Availability and implementation The source code, pretrained and finetuned model for DNABERT are available at GitHub (https://github.com/jerryji1993/DNABERT). Supplementary information Supplementary data are available at Bioinformatics online.

Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Dalla-Torre et al. (2024). *Nature Methods*. doi:10.1038/s41592-024-02523-z

> The prediction of molecular phenotypes from DNA sequences remains a longstanding challenge in genomics, often driven by limited annotated data and the inability to transfer learnings between tasks. Here, we present an extensive study of foundation models pre-trained on DNA sequences, named Nucleotide Transformer, ranging from 50 million up to 2.5 billion parameters and integrating information from 3,202 human genomes and 850 genomes from diverse species. These transformer models yield context-specific representations of nucleotide sequences, which allow for accurate predictions even in low-data settings. We show that the developed models can be fine-tuned at low cost to solve a variety of genomics applications. Despite no supervision, the models learned to focus attention on key genomic elements and can be used to improve the prioritization of genetic variants. The training and application of foundational models in genomics provides a widely applicable approach for accurate molecular phenotype prediction from DNA sequence.

- Nguyen et al. (2024), Brixi et al. (2025) — Evo and Evo 2.

Sequence modeling and design from molecular to genome scale with Evo. Nguyen et al. (2024). *Science*. doi:10.1126/science.ado9336

> The genome is a sequence that encodes the DNA, RNA, and proteins that orchestrate an organism's function. We present Evo, a long-context genomic foundation model with a frontier architecture trained on millions of prokaryotic and phage genomes, and report scaling laws on DNA to complement observations in language and vision. Evo generalizes across DNA, RNA, and proteins, enabling zero-shot function prediction competitive with domain-specific language models and the generation of functional CRISPR-Cas and transposon systems, representing the first examples of protein-RNA and protein-DNA codesign with a language model. Evo also learns how small mutations affect whole-organism fitness and generates megabase-scale sequences with plausible genomic architecture. These prediction and generation capabilities span molecular to genomic scales of complexity, advancing our understanding and control of biology.

Genome modeling and design across all domains of life with Evo 2. Brixi et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.02.18.638918

> All of life encodes information with DNA. While tools for sequencing, synthesis, and editing of genomic code have transformed biological research, intelligently composing new biological systems would also require a deep understanding of the immense complexity encoded by genomes. We introduce Evo 2, a biological foundation model trained on 9.3 trillion DNA base pairs from a highly curated genomic atlas spanning all domains of life. We train Evo 2 with 7B and 40B parameters to have an unprecedented 1 million token context window with single-nucleotide resolution. Evo 2 learns from DNA sequence alone to accurately predict the functional impacts of genetic variation—from noncoding pathogenic mutations to clinically significant BRCA1 variants—without task-specific finetuning. Applying mechanistic interpretability analyses, we reveal that Evo 2 autonomously learns a breadth of biological features, including exon–intron boundaries, transcription factor binding sites, protein structural elements, and prophage genomic regions. Beyond its predictive capabilities, Evo 2 generates mitochondrial, prokaryotic, and eukaryotic sequences at genome scale with greater naturalness and coherence than previous methods. Guiding Evo 2 via inference-time search enables controllable generation of epigenomic structure, for which we demonstrate the first inference-time scaling results in biology. We make Evo 2 fully open, including model parameters, training code, inference code, and the OpenGenome2 dataset, to accelerate the exploration and design of biological complexity.

- Avsec et al. (2021), Avsec et al. (2026) — Enformer and AlphaGenome.

Effective gene expression prediction from sequence by integrating long-range interactions. Avsec et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01252-x

> How noncoding DNA determines gene expression in different cell types is a major unsolved problem, and critical downstream applications in human genetics depend on improved solutions. Here, we report substantially improved gene expression prediction accuracy from DNA sequences through the use of a deep learning architecture, called Enformer, that is able to integrate information from long-range interactions (up to 100 kb away) in the genome. This improvement yielded more accurate variant effect predictions on gene expression for both natural genetic variants and saturation mutagenesis measured by massively parallel reporter assays. Furthermore, Enformer learned to predict enhancer-promoter interactions directly from the DNA sequence competitively with methods that take direct experimental data as input. We expect that these advances will enable more effective fine-mapping of human disease associations and provide a framework to interpret cis-regulatory evolution.

Advancing regulatory variant effect prediction with AlphaGenome. Avsec et al. (2026). *Nature*. doi:10.1038/s41586-025-10014-0

> Deep learning models that predict functional genomic measurements from DNA sequences are powerful tools for deciphering the genetic regulatory code. Existing methods involve a trade-off between input sequence length and prediction resolution, thereby limiting their modality scope and performance 1-5 . We present AlphaGenome, a unified DNA sequence model, which takes as input 1 Mb of DNA sequence and predicts thousands of functional genomic tracks up to single-base-pair resolution across diverse modalities. The modalities include gene expression, transcription initiation, chromatin accessibility, histone modifications, transcription factor binding, chromatin contact maps, splice site usage and splice junction coordinates and strength. Trained on human and mouse genomes, AlphaGenome matches or exceeds the strongest available external models in 25 of 26 evaluations of variant effect prediction. The ability of AlphaGenome to simultaneously score variant effects across all modalities accurately recapitulates the mechanisms of clinically relevant variants near the TAL1 oncogene 6 . To facilitate broader use, we provide tools for making genome track and variant effect predictions from sequence.

- Novakovsky et al. (2022) — explainable AI in genetics.

Obtaining genetics insights from deep learning via explainable artificial intelligence. Novakovsky et al. (2022). *Nature Reviews Genetics*. doi:10.1038/s41576-022-00532-2

> Artificial intelligence (AI) models based on deep learning now represent the state of the art for making functional predictions in genomics research. However, the underlying basis on which predictive models make such predictions is often unknown. For genomics researchers, this missing explanatory information would frequently be of greater value than the predictions themselves, as it can enable new insights into genetic processes. We review progress in the emerging area of explainable AI (xAI), a field with the potential to empower life science researchers to gain mechanistic insights into complex deep learning models. We discuss and categorize approaches for model interpretation, including an intuitive understanding of how each approach works and their underlying assumptions and limitations in the context of typical high-throughput biological datasets.

B.12.6 Generative and foundation models

- Kingma and Welling (2014), Goodfellow et al. (2020), Ho et al. (2020) — VAEs, GANs and diffusion models.

  Auto-Encoding Variational Bayes. Kingma & Welling (2014). https://arxiv.org/abs/1312.6114

  Generative adversarial networks. Goodfellow et al. (2020). *Communications of the ACM*. doi:10.1145/3422622

  > Generative adversarial networks are a kind of artificial intelligence algorithm designed to solve the generative modeling problem. The goal of a generative model is to study a collection of training examples and learn the probability distribution that generated them. Generative Adversarial Networks (GANs) are then able to generate more examples from the estimated probability distribution. Generative models based on deep learning are common, but GANs are among the most successful generative models (especially in terms of their ability to generate realistic high-resolution images). GANs have been successfully applied to a wide variety of tasks (mostly in research settings) but continue to present unique challenges and research opportunities because they are based on game theory while most other approaches to generative modeling are based on optimization.

  Denoising Diffusion Probabilistic Models. Ho et al. (2020). *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2006.11239

- Madani et al. (2023), Ferruz et al. (2022) — generating protein sequences with language models.

  Large language models generate functional protein sequences across diverse families. Madani et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-022-01618-2

  > Deep-learning language models have shown promise in various biotechnological applications, including protein design and engineering. Here we describe ProGen, a language model that can generate protein sequences with a predictable function across large protein families, akin to generating grammatically and semantically correct natural language sentences on diverse topics. The model was trained on 280 million protein sequences from >19,000 families and is augmented with control tags specifying protein properties. ProGen can be further fine-tuned to curated sequences and tags to improve controllable generation performance of proteins from families with sufficient homologous samples. Artificial proteins fine-tuned to five distinct lysozyme families showed similar catalytic efficiencies as natural lysozymes, with sequence identity to natural proteins as low as 31.4%. ProGen is readily adapted to diverse protein families, as we demonstrate with chorismate mutase and malate dehydrogenase.

  ProtGPT2 is a deep unsupervised language model for protein design. Ferruz et al. (2022). *Nature Communications*. doi:10.1038/s41467-022-32007-7

  > Protein design aims to build novel proteins customized for specific purposes, thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in Transformer-based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Here, motivated by this success, we describe ProtGPT2, a language model trained on the protein space that generates de novo protein sequences following the principles of natural ones. The generated proteins display natural amino acid propensities, while disorder predictions indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are distantly related to natural ones, and similarity networks further demonstrate that ProtGPT2 is sampling unexplored regions of protein space. AlphaFold prediction of ProtGPT2-sequences yields well-folded non-idealized structures with embodiments and large loops and reveals topologies not captured in current structure databases. ProtGPT2 generates sequences in a matter of seconds and is freely available.

- Hayes et al. (2025) — ESM3.

  Simulating 500 million years of evolution with a language model. Hayes et al. (2025). *Science*. doi:10.1126/science.ads0018

> More than 3 billion years of evolution have produced an image of biology encoded into the space of natural proteins. Here, we show that language models trained at scale on evolutionary data can generate functional proteins that are far away from known proteins. We present ESM3, a frontier multimodal generative language model that reasons over the sequence, structure, and function of proteins. ESM3 can follow complex prompts combining its modalities and is highly responsive to alignment to improve its fidelity. We have prompted ESM3 to generate fluorescent proteins. Among the generations that we synthesized, we found a bright fluorescent protein at a far distance (58% sequence identity) from known fluorescent proteins, which we estimate is equivalent to simulating 500 million years of evolution.

- Bommasani et al. (2021) — foundation models.

  On the Opportunities and Risks of Foundation Models. Bommasani & others (2021). https://arxiv.org/abs/2108.07258

- Kaplan et al. (2020) — scaling laws.

  Scaling Laws for Neural Language Models. Kaplan et al. (2020). https://arxiv.org/abs/2001.08361

- Thorp (2023) — ChatGPT is fun, but not an author.

  ChatGPT is fun, but not an author. Thorp (2023). *Science*. doi:10.1126/science.adg7879

  > In less than 2 months, the artificial intelligence (AI) program ChatGPT has become a cultural sensation. It is freely accessible through a web portal created by the tool's developer, OpenAI. The program-which automatically creates text based on written prompts-is so popular that it's likely to be "at capacity right now" if you attempt to use it. When you do get through, ChatGPT provides endless entertainment. I asked it to rewrite the first scene of the classic American play Death of a Salesman , but to feature Princess Elsa from the animated movie Frozen as the main character instead of Willy Loman. The output was an amusing conversation in which Elsa-who has come home from a tough day of selling-is told by her son Happy, "Come on, Mom. You're Elsa from Frozen . You have ice powers and you're a queen. You're unstoppable." Mash-ups like this are certainly fun, but there are serious implications for generative AI programs like ChatGPT in science and academia.

- James et al. (2021) — *An Introduction to Statistical Learning* (free online).

  An Introduction to Statistical Learning. James et al. (2021). *Springer*. https://www.statlearning.com/

- Notin et al. (2023) — ProteinGym.

  ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. Notin et al. (2023). *Advances in Neural Information Processing Systems*. https://proteingym.org

- Kedzierska et al. (2025), Ahlmann-Eltze et al. (2025) — evaluating foundation models against simple baselines.

  Zero-shot evaluation reveals limitations of single-cell foundation models. Kedzierska et al. (2025). *Genome Biology*. doi:10.1186/s13059-025-03574-x

  > Foundation models such as scGPT and Geneformer have not been rigorously evaluated in a setting where they are used without any further training (i.e., zero-shot). Understanding the performance of models in zero-shot settings is critical to applications that exclude the ability to fine-tune, such as discovery settings where labels are unknown. Our evaluation of the zero-shot performance of Geneformer and scGPT suggests that, in some cases, these models may face reliability challenges and could be outperformed by simpler methods. Our findings underscore the importance of zero-shot evaluations in development and deployment of foundation models in single-cell research.

  Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Ahlmann-Eltze et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02772-6

Recent research in deep-learning-based foundation models promises to learn representations of single-cell data that enable prediction of the effects of genetic perturbations. Here we compared five foundation models and two other deep learning models against deliberately simple baselines for predicting transcriptome changes after single or double perturbations. None outperformed the baselines, which highlights the importance of critical benchmarking in directing and evaluating method development.

- Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Žídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, et al. (2021) — AlphaFold2 (see Chapter 9).

  Highly accurate protein structure prediction with AlphaFold. Jumper et al. (2021). *Nature*. doi:10.1038/s41586-021-03819-2

  Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort 1-4 , the structures of around 100,000 unique proteins have been determined 5 , but this represents a small fraction of the billions of known protein sequences 6,7 . Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence-the structure prediction component of the 'protein folding problem' 8 -has been an important open research problem for more than 50 years 9 . Despite recent progress 10-14 , existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14) 15 , demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

B.13 Image analysis and models of cells

Further papers for Chapter 13.

In the HTML version, click *Abstract* to show the abstract.

B.13.1 Reviews and introductions

- Moen et al. (2019) — deep learning for cellular image analysis.

  Deep learning for cellular image analysis. Moen et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0403-1

  Recent advances in computer vision and machine learning underpin a collection of algorithms with an impressive ability to decipher the content of images. These deep learning algorithms are being applied to biological images and are transforming the analysis and interpretation of imaging data. These advances are positioned to render difficult analyses routine and to enable researchers to carry out new, previously impossible experiments. Here we review the intersection between deep learning and cellular image analysis and provide an overview of both the mathematical mechanics and the programming frameworks of deep learning that are pertinent to life scientists. We survey the field's progress in four key applications: image classification, image segmentation, object tracking, and augmented microscopy. Last, we relay our labs' experience with three key aspects of implementing deep learning in the laboratory: annotating training data, selecting and training a range of neural network architectures, and deploying solutions. We also highlight existing datasets and implementations for each surveyed application.

- Hallou et al. (2021) — deep learning for bioimage analysis, written for biologists.

  Deep learning for bioimage analysis in developmental biology. Hallou et al. (2021). *Development*. doi:10.1242/dev.199616

  > Deep learning has transformed the way large and complex image datasets can be processed, reshaping what is possible in bioimage analysis. As the complexity and size of bioimage data continues to grow, this new analysis paradigm is becoming increasingly ubiquitous. In this Review, we begin by introducing the concepts needed for beginners to understand deep learning. We then review how deep learning has impacted bioimage analysis and explore the open-source resources available to integrate it into a research project. Finally, we discuss the future of deep learning applied to cell and developmental biology. We analyze how state-of-the-art methodologies have the potential to transform our understanding of biological systems through new image-based analysis and modelling that integrate multimodal inputs in space and time.

- Laine et al. (2021) — avoiding a replication crisis in deep-learning image analysis.

  Avoiding a replication crisis in deep-learning-based bioimage analysis. Laine et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01284-3

- Lundberg and Borner (2019) — spatial proteomics as a tool for cell biology.

  Spatial proteomics: a powerful discovery tool for cell biology. Lundberg & Borner (2019). *Nature Reviews Molecular Cell Biology*. doi:10.1038/s41580-018-0094-y

  > Protein subcellular localization is tightly controlled and intimately linked to protein function in health and disease. Capturing the spatial proteome - that is, the localizations of proteins and their dynamics at the subcellular level - is therefore essential for a complete understanding of cell biology. Owing to substantial advances in microscopy, mass spectrometry and machine learning applications for data analysis, the field is now mature for proteome-wide investigations of spatial cellular regulation. Studies of the human proteome have begun to reveal a complex architecture, including single-cell variations, dynamic protein translocations, changing interaction networks and proteins localizing to multiple compartments. Furthermore, several studies have successfully harnessed the power of comparative spatial proteomics as a discovery tool to unravel disease mechanisms. We are at the beginning of an era in which spatial proteomics finally integrates with cell biology and medical research, thereby paving the way for unbiased systems-level insights into cellular processes. Here, we discuss current methods for spatial proteomics using imaging or mass spectrometry and specifically highlight global comparative applications. The aim of this Review is to survey the state of the field and also to encourage more cell biologists to apply spatial proteomics approaches.

B.13.2 Deep learning for images

- Krizhevsky et al. (2017), He et al. (2016) — AlexNet and residual networks.

  ImageNet classification with deep convolutional neural networks. Krizhevsky et al. (2017). *Communications of the ACM*. doi:10.1145/3065386

  > We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully connected layers we employed a recently developed regularization method called ”dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

  Deep Residual Learning for Image Recognition. He et al. (2016). *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.90

> Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers - 8× deeper than VGG nets [40] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions1, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

- Ronneberger et al. (2015), Falk et al. (2018) — U-Net, and U-Net for cell biologists.

  U-Net: Convolutional Networks for Biomedical Image Segmentation. Ronneberger et al. (2015). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. doi:10.1007/978-3-319-24574-4_28

  U-Net: deep learning for cell counting, detection, and morphometry. Falk et al. (2018). *Nature Methods*. doi:10.1038/s41592-018-0261-2

  > U-Net is a generic deep-learning solution for frequently occurring quantification tasks such as cell detection and shape measurements in biomedical image data. We present an ImageJ plugin that enables non-machine-learning experts to analyze their data with U-Net on either a local computer or a remote server/cloud service. The plugin comes with pretrained models for single-cell segmentation and allows for U-Net to be adapted to new tasks on the basis of a few annotated samples.

- Dosovitskiy et al. (2021) — the vision transformer.

  An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. (2021). doi:10.48550/arXiv.2010.11929

  > While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

- Caron et al. (2021), Oquab et al. (2023), He et al. (2022) — self-supervised learning of images (DINO, DINOv2, masked autoencoders).

  Emerging Properties in Self-Supervised Vision Transformers. Caron et al. (2021). *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/iccv48922.2021.00951

  > In this paper, we question if self-supervised learning provides new properties to Vision Transformer (ViT) [16] that stand out compared to convolutional networks (convnets). Beyond the fact that adapting self-supervised methods to this architecture works particularly well, we make the following observations: first, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent k-NN classifiers, reaching 78.3% top-1 on ImageNet with a small ViT. Our study also underlines the importance of momentum encoder [26], multi-crop training [9], and the use of small patches with ViTs. We implement our findings into a simple self-supervised method, called DINO, which we interpret as a form of self-distillation with no labels. We show the synergy between DINO and ViTs by achieving 80.1% top-1 on ImageNet in linear evaluation with ViT-Base.

DINOv2: Learning Robust Visual Features without Supervision. Oquab et al. (2023). doi:10.48550/arXiv.2304.07193

> The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar foundation models in computer vision. These models could greatly simplify the use of images in any system by producing all-purpose visual features, i.e., features that work across image distributions and tasks without finetuning. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit existing approaches and combine different techniques to scale our pretraining in terms of data and model size. Most of the technical contributions aim at accelerating and stabilizing the training at scale. In terms of data, we propose an automatic pipeline to build a dedicated, diverse, and curated image dataset instead of uncurated data, as typically done in the self-supervised literature. In terms of models, we train a ViT model (Dosovitskiy et al., 2020) with 1B parameters and distill it into a series of smaller models that surpass the best available all-purpose features, OpenCLIP (Ilharco et al., 2021) on most of the benchmarks at image and pixel levels.

Masked Autoencoders Are Scalable Vision Learners. He et al. (2022). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr52688.2022.01553

> This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3× or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. Transfer performance in downstream tasks outperforms supervised pretraining and shows promising scaling behavior.

- Kirillov et al. (2023), Archit et al. (2025), Marks et al. (2025) — Segment Anything and its versions for microscopy.

Segment Anything. Kirillov et al. (2023). *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/iccv51070.2023.00371

> We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 billion masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive – often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at segment-anything.com to foster research into foundation models for computer vision. We recommend reading the full paper at: arxiv.org/abs/2304.02643.

Segment Anything for Microscopy. Archit et al. (2025). *Nature Methods*. doi:10.1038/s41592-024-02580-4

> Accurate segmentation of objects in microscopy images remains a bottleneck for many researchers despite the number of tools developed for this purpose. Here, we present Segment Anything for Microscopy (μSAM), a tool for segmentation and tracking in multidimensional microscopy data. It is based on Segment Anything, a vision foundation model for image segmentation. We extend it by fine-tuning generalist models for light and electron microscopy that clearly improve segmentation quality for a wide range of imaging conditions. We also implement interactive and automatic segmentation in a napari plugin that can speed up diverse segmentation tasks and provides a unified solution for microscopy annotation across different microscopy modalities. Our work constitutes the application of vision foundation models in microscopy, laying the groundwork for solving image analysis tasks in this domain with a small set of powerful deep learning models.

CellSAM: a foundation model for cell segmentation. Marks et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02879-w

> Cells are a fundamental unit of biological organization, and identifying them in imaging data-cell segmentation-is a critical task for various cellular imaging experiments. Although deep learning methods have led to substantial progress on this problem, most models are specialist models that work well for specific domains but cannot be applied across domains or scale well with large amounts of data. Here we present CellSAM, a universal model for cell segmentation that generalizes across diverse cellular imaging data. CellSAM builds on top of the Segment Anything Model (SAM) by developing a prompt engineering approach for mask generation. We train an object detector, CellFinder, to automatically detect cells and prompt SAM to generate segmentations. We show that this approach allows a single model to achieve human-level performance for segmenting images of mammalian cells, yeast and bacteria collected across various imaging modalities. We show that CellSAM has strong zero-shot performance and can be improved with a few examples via few-shot learning. Additionally, we demonstrate how CellSAM can be applied across diverse bioimage analysis workflows. A deployed version of CellSAM is available at https://cellsam.deepcell.org/ .

- Moutakanni et al. (2025) — self-supervised features for fluorescence microscopy (Cell-DINO).

  Cell-DINO: Self-supervised image-based embeddings for cell fluorescent microscopy. Moutakanni et al. (2025). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1013828

  > Accurately quantifying cellular morphology at scale could substantially empower existing single-cell approaches. However, measuring cell morphology remains an active field of research, which has inspired multiple computer vision algorithms over the years. Here, we show that DINOv2, a vision-transformer based, self-supervised algorithm, has a remarkable ability for learning rich representations of cellular morphology without manual annotations or any other type of supervision. We apply DINOv2 to cell phenotyping problems, and compare the performance of resulting models, called Cell-DINO models, on a wide variety of tasks across two publicly available imaging datasets of diverse specifications and biological focus. Compared to supervised and other self-supervised baselines, Cell-DINO models demonstrate improved performance, especially in low annotation regimes. For instance, to classify protein localization using only 1% of annotations on a challenging single-cell dataset, Cell-DINO performs 70% better than a supervised strategy, and 24% better than another self-supervised alternative. The results show that Cell-DINO can support the study of unknown biological variation, including single-cell heterogeneity and relationships between experimental conditions, making it an excellent tool for image-based biological discovery.

B.13.3 Segmentation and label-free prediction

- Schmidt et al. (2018) — StarDist.

  Cell Detection with Star-Convex Polygons. Schmidt et al. (2018). *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. doi:10.1007/978-3-030-00934-2_30

- Pachitariu and Stringer (2022) — Cellpose 2.0 and human-in-the-loop training.

  Cellpose 2.0: how to train your own model. Pachitariu & Stringer (2022). *Nature Methods*. doi:10.1038/s41592-022-01663-4

  > Pretrained neural network models for biological segmentation can provide good out-of-the-box results for many image types. However, such models do not allow users to adapt the segmentation style to their specific needs and can perform suboptimally for test images that are very different from the training images. Here we introduce Cellpose 2.0, a new package that includes an ensemble of diverse pretrained models as well as a human-in-the-loop pipeline for rapid prototyping of new custom models. We show that models pretrained on the Cellpose dataset can be fine-tuned with only 500-1,000 user-annotated regions of interest (ROI) to perform nearly as well as models trained on entire datasets with up to 200,000 ROI. A human-in-the-loop approach further reduced the required user annotation to 100-200 ROI, while maintaining high-quality segmentations. We provide software tools such as an annotation graphical user interface, a model zoo and a human-in-the-loop pipeline to facilitate the adoption of Cellpose 2.0.

- Greenwald et al. (2021) — whole-cell segmentation of tissue images (Mesmer, TissueNet).

  Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Greenwald et al. (2021). *Nature Biotechnology*. doi:10.1038/s41587-021-01094-0

  > A principal challenge in the analysis of tissue imaging data is cell segmentation-the task of identifying the precise boundary of every cell in an image. To address this problem we constructed TissueNet, a dataset for training segmentation models that contains more than 1 million manually labeled cells, an order of magnitude more than all previously published segmentation training datasets. We used TissueNet to train Mesmer, a deep-learning-enabled segmentation algorithm. We demonstrated that Mesmer is more accurate than previous methods, generalizes to the full diversity of tissue types and imaging platforms in TissueNet, and achieves human-level performance. Mesmer enabled the automated extraction of key cellular features, such as subcellular localization of protein signal, which was challenging with previous approaches. We then adapted Mesmer to harness cell lineage information in highly multiplexed datasets and used this enhanced version to quantify cell morphology changes during human gestation. All code, data and models are released as a community resource.

- Christiansen et al. (2018), Ounkomol et al. (2018) — predicting fluorescence images from transmitted light.

  In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. Christiansen et al. (2018). *Cell*. doi:10.1016/j.cell.2018.03.040

  > Microscopy is a central method in life sciences. Many popular methods, such as antibody labeling, are used to add physical fluorescent labels to specific cellular constituents. However, these approaches have significant drawbacks, including inconsistency; limitations in the number of simultaneous labels because of spectral overlap; and necessary perturbations of the experiment, such as fixing the cells, to generate the measurement. Here, we show that a computational machine-learning approach, which we call "in silico labeling" (ISL), reliably predicts some fluorescent labels from transmitted-light images of unlabeled fixed or live biological samples. ISL predicts a range of labels, such as those for nuclei, cell type (e.g., neural), and cell state (e.g., cell death). Because prediction happens in silico, the method is consistent, is not limited by spectral overlap, and does not disturb the experiment. ISL generates biological measurements that would otherwise be problematic or impossible to acquire.

  Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Ounkomol et al. (2018). *Nature Methods*. doi:10.1038/s41592-018-0111-2

  > Understanding cells as integrated systems is central to modern biology. Although fluorescence microscopy can resolve subcellular structure in living cells, it is expensive, is slow, and can damage cells. We present a label-free method for predicting three-dimensional fluorescence directly from transmitted-light images and demonstrate that it can be used to generate multi-structure, integrated images. The method can also predict immunofluorescence (IF) from electron micrograph (EM) inputs, extending the potential applications.

B.13.4 Image-based profiling

- Gustafsdottir et al. (2013), M.-A. Bray et al. (2016), Cimini et al. (2023) — the Cell Painting assay and its optimisation.

  Multiplex Cytological Profiling Assay to Measure Diverse Cellular States. Gustafsdottir et al. (2013). *PLoS ONE*. doi:10.1371/journal.pone.0080999

  > Computational methods for image-based profiling are under active development, but their success hinges on assays that can capture a wide range of phenotypes. We have developed a multiplex cytological profiling assay that "paints the cell" with as many fluorescent markers as possible without compromising our ability to extract rich, quantitative profiles in high throughput. The assay detects seven major cellular components. In a pilot screen of bioactive compounds, the assay detected a range of cellular phenotypes and it clustered compounds with similar annotated protein targets or chemical structure based on cytological profiles. The results demonstrate that the assay captures subtle patterns in the combination of morphological labels, thereby detecting the effects of chemical compounds even though their targets

are not stained directly. This image-based assay provides an unbiased approach to characterize compound- and disease-associated cell states to support future probe discovery.

Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Bray et al. (2016). *Nature Protocols*. doi:10.1038/nprot.2016.105

> In morphological profiling, quantitative data are extracted from microscopy images of cells to identify biologically relevant similarities and differences among samples based on these profiles. This protocol describes the design and execution of experiments using Cell Painting, which is a morphological profiling assay that multiplexes six fluorescent dyes, imaged in five channels, to reveal eight broadly relevant cellular components or organelles. Cells are plated in multiwell plates, perturbed with the treatments to be tested, stained, fixed, and imaged on a high-throughput microscope. Next, an automated image analysis software identifies individual cells and measures ~1,500 morphological features (various measures of size, shape, texture, intensity, and so on) to produce a rich profile that is suitable for the detection of subtle phenotypes. Profiles of cell populations treated with different experimental perturbations can be compared to suit many goals, such as identifying the phenotypic impact of chemical or genetic perturbations, grouping compounds and/or genes into functional pathways, and identifying signatures of disease. Cell culture and image acquisition takes 2 weeks; feature extraction and data analysis take an additional 1-2 weeks.

Optimizing the Cell Painting assay for image-based profiling. Cimini et al. (2023). *Nature Protocols*. doi:10.1038/s41596-023-00840-9

> In image-based profiling, software extracts thousands of morphological features of cells from multi-channel fluorescence microscopy images, yielding single-cell profiles that can be used for basic research and drug discovery. Powerful applications have been proven, including clustering chemical and genetic perturbations on the basis of their similar morphological impact, identifying disease phenotypes by observing differences in profiles between healthy and diseased cells and predicting assay outcomes by using machine learning, among many others. Here, we provide an updated protocol for the most popular assay for image-based profiling, Cell Painting. Introduced in 2013, it uses six stains imaged in five channels and labels eight diverse components of the cell: DNA, cytoplasmic RNA, nucleoli, actin, Golgi apparatus, plasma membrane, endoplasmic reticulum and mitochondria. The original protocol was updated in 2016 on the basis of several years' experience running it at two sites, after optimizing it by visual stain quality. Here, we describe the work of the Joint Undertaking for Morphological Profiling Cell Painting Consortium, to improve upon the assay via quantitative optimization by measuring the assay's ability to detect morphological phenotypes and group similar perturbations together. The assay gives very robust outputs despite various changes to the protocol, and two vendors' dyes work equivalently well. We present Cell Painting version 3, in which some steps are simplified and several stain concentrations can be reduced, saving costs. Cell culture and image acquisition take 1-2 weeks for typically sized batches of ≤20 plates; feature extraction and data analysis take an additional 1-2 weeks.This protocol is an update to Nat. Protoc. 11, 1757-1774 (2016): https://doi.org/10.1038/nprot.2016.105.

- Caicedo et al. (2017) — data-analysis strategies for image-based profiling.

  Data-analysis strategies for image-based cell profiling. Caicedo et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4397

  > Image-based cell profiling is a high-throughput strategy for the quantification of phenotypic differences among a variety of cell populations. It paves the way to studying biological systems on a large scale by using chemical and genetic perturbations. The general workflow for this technology involves image acquisition with high-throughput microscopy systems and subsequent image processing and analysis. Here, we introduce the steps required to create high-quality image-based (i.e., morphological) profiles from a collection of microscopy images. We recommend techniques that have proven useful in each stage of the data analysis process, on the basis of the experience of 20 laboratories worldwide that are refining their image-based cell-profiling methodologies in pursuit of biological discovery. The recommended techniques cover alternatives that may suit various biological goals, experimental designs, and laboratories' preferences.

- Chandrasekaran et al. (2025) — a morphological map of gene over- and underexpression.

  Morphological map of under- and overexpression of genes in human cells. Chandrasekaran et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02753-9

> Cell Painting images offer valuable insights into a cell's state and enable many biological applications, but publicly available arrayed datasets only include hundreds of genes perturbed. The JUMP Cell Painting Consortium perturbed roughly 75% of the protein-coding genome in human U-2 OS cells, generating a rich resource of single-cell images and extracted features. These profiles capture the phenotypic impacts of perturbing 15,243 human genes, including overexpressing 12,609 genes (using open reading frames) and knocking out 7,975 genes (using CRISPR-Cas9). Here we mitigated technical artifacts by rigorously evaluating data processing options and validated the dataset's robustness and biological relevance. Analysis of phenotypic profiles revealed previously undiscovered gene clusters and functional relationships, including those associated with mitochondrial function, cancer and neural processes. The JUMP Cell Painting genetic dataset is a valuable resource for exploring gene relationships and uncovering previously unknown functions.

- Fay et al. (2023) — RxRx3, a public phenomics map (preprint).

  RxRx3: Phenomics Map of Biology. Fay et al. (2023). *bioRxiv (preprint)*. doi:10.1101/2023.02.07.527350

  > The combination of modern genetic perturbation techniques with high content screening has enabled genome-scale cell microscopy experiments that can be leveraged to construct maps of biology . These are built by processing microscopy images to produce readouts in unified and relatable representation space to capture known biological relationships and discover new ones. To further enable the scientific community to develop methods and insights from map-scale data, here we release RxRx3 , the first ever public high-content screening dataset combining genome-scale CRISPR knockouts with multiple-concentration screening of small molecules (a set of FDA approved and commercially available bioactive compounds). The dataset contains 6-channel fluorescent microscopy images and associated deep learning embeddings from over 2.2 million wells that span 17,063 CRISPR knockouts and 1,674 compounds at 8 doses each. RxRx3 is one of the largest collections of cellular screening data, and as far as we know, the largest generated consistently via a common experimental protocol within a single laboratory. Our goal in releasing RxRx3 is to demonstrate the benefits of generating consistent data, enable the development of the machine learning methods on this scale of data and to foster research, methods development, and collaboration. For more information about RxRx3 please visit RxRx.ai/rxrx3

B.13.5 Protein localisation

- Thul et al. (2017), Mahdessian et al. (2021) — the subcellular map of the Human Protein Atlas and single-cell variation.

  A subcellular map of the human proteome. Thul et al. (2017). *Science*. doi:10.1126/science.aal3321

  > Resolving the spatial distribution of the human proteome at a subcellular level can greatly increase our understanding of human biology and disease. Here we present a comprehensive image-based map of subcellular protein distribution, the Cell Atlas, built by integrating transcriptomics and antibody-based immunofluorescence microscopy with validation by mass spectrometry. Mapping the in situ localization of 12,003 human proteins at a single-cell level to 30 subcellular structures enabled the definition of the proteomes of 13 major organelles. Exploration of the proteomes revealed single-cell variations in abundance or spatial distribution and localization of about half of the proteins to multiple compartments. This subcellular map can be used to refine existing protein-protein interaction networks and provides an important resource to deconvolute the highly complex architecture of the human cell.

  Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Mahdessian et al. (2021). *Nature*. doi:10.1038/s41586-021-03232-9

  > The cell cycle, over which cells grow and divide, is a fundamental process of life. Its dysregulation has devastating consequences, including cancer 1-3 . The cell cycle is driven by precise regulation of proteins in time and space, which creates variability between individual proliferating cells. To our knowledge, no systematic investigations of such cell-to-cell proteomic variability exist. Here we present a comprehensive, spatiotemporal map of human proteomic heterogeneity by integrating proteomics at subcellular resolution with single-cell transcriptomics and precise temporal measurements of individual cells in the cell cycle. We show that around one-fifth of the human proteome displays cell-to-cell variability, identify hundreds of proteins with previously unknown associations with mitosis and the cell cycle, and provide evidence that several of these proteins have oncogenic functions. Our results show that cell cycle progression explains

> less than half of all cell-to-cell variability, and that most cycling proteins are regulated post-translationally, rather than by transcriptomic cycling. These proteins are disproportionately phosphorylated by kinases that regulate cell fate, whereas non-cycling proteins that vary between cells are more likely to be modified by kinases that regulate metabolism. This spatially resolved proteomic map of the cell cycle is integrated into the Human Protein Atlas and will serve as a resource for accelerating molecular studies of the human cell cycle and cell proliferation.

- Le et al. (2022) — the weakly supervised single-cell classification competition.

  Analysis of the Human Protein Atlas Weakly Supervised Single-Cell Classification competition. Le et al. (2022). *Nature Methods*. doi:10.1038/s41592-022-01606-z

  > While spatial proteomics by fluorescence imaging has quickly become an essential discovery tool for researchers, fast and scalable methods to classify and embed single-cell protein distributions in such images are lacking. Here, we present the design and analysis of the results from the competition Human Protein Atlas - Single-Cell Classification hosted on the Kaggle platform. This represents a crowd-sourced competition to develop machine learning models trained on limited annotations to label single-cell protein patterns in fluorescent images. The particular challenges of this competition include class imbalance, weak labels and multi-label classification, prompting competitors to apply a wide range of approaches in their solutions. The winning models serve as the first subcellular omics tools that can annotate single-cell locations, extract single-cell features and capture cellular dynamics.

- Cho et al. (2022), Kobayashi et al. (2022) — OpenCell and self-supervised localisation features (cytoself).

  OpenCell: Endogenous tagging for the cartography of human cellular organization. Cho et al. (2022). *Science*. doi:10.1126/science.abi6983

  > Elucidating the wiring diagram of the human cell is a central goal of the postgenomic era. We combined genome engineering, confocal live-cell imaging, mass spectrometry, and data science to systematically map the localization and interactions of human proteins. Our approach provides a data-driven description of the molecular and spatial networks that organize the proteome. Unsupervised clustering of these networks delineates functional communities that facilitate biological discovery. We found that remarkably precise functional information can be derived from protein localization patterns, which often contain enough information to identify molecular interactions, and that RNA binding proteins form a specific subgroup defined by unique interaction and localization properties. Paired with a fully interactive website (opencell.czbiohub.org), our work constitutes a resource for the quantitative cartography of human cellular organization.

  Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Kobayashi et al. (2022). *Nature Methods*. doi:10.1038/s41592-022-01541-z

  > Explaining the diversity and complexity of protein localization is essential to fully understand cellular architecture. Here we present cytoself, a deep-learning approach for fully self-supervised protein localization profiling and clustering. Cytoself leverages a self-supervised training scheme that does not require preexisting knowledge, categories or annotations. Training cytoself on images of 1,311 endogenously labeled proteins from the OpenCell database reveals a highly resolved protein localization atlas that recapitulates major scales of cellular organization, from coarse classes, such as nuclear and cytoplasmic, to the subtle localization signatures of individual protein complexes. We quantitatively validate cytoself's ability to cluster proteins into organelles and protein complexes, showing that cytoself outperforms previous self-supervised approaches. Moreover, to better understand the inner workings of our model, we dissect the emergent features from which our clustering is derived, interpret them in the context of the fluorescence images, and analyze the performance contributions of each component of our approach.

- Gupta et al. (2024) — SubCell, vision foundation models of the HPA images (preprint).

  SubCell: Proteome-aware vision foundation models for microscopy capture single-cell biology. Gupta et al. (2024). *bioRxiv (preprint)*. doi:10.1101/2024.12.06.627299

> Cell morphology and subcellular protein organization provide important insights into cellular function and behavior. These cellular features can be studied using large-scale fluorescence microscopy, and machine learning has become a powerful tool to interpret the resulting images for biological insights. Here, we introduce SubCell, a deep learning model for fluorescence microscopy designed to accurately capture cellular morphology, protein localization, cellular organization, and biological function beyond what humans can readily perceive. SubCell was trained on the proteome-wide image collection from the Human Protein Atlas with a novel proteome-aware learning objective. SubCell outperforms state-of-the-art methods across a variety of tasks relevant to single-cell biology and generalizes to other fluorescence microscopy datasets without any fine-tuning. Additionally, we construct the first proteome-wide hierarchical map of proteome organization that is directly learned from image data. This vision-based multiscale cell map defines cellular subsystems down to protein complex resolution, reveals proteins with similar functions, and distinguishes dynamic and stable behaviors within cellular compartments. Finally, combining SubCell with a protein sequence model enables a rich multimodal approach to capture gene function better than either vision-only or sequence-only models alone. In conclusion, SubCell creates deep, image-driven representations of cellular architecture that are applicable across diverse biological contexts and datasets.

- Xiao et al. (2024) — review of AI methods for protein localisation.

  A Review for Artificial Intelligence Based Protein Subcellular Localization. Xiao et al. (2024). *Biomolecules*. doi:10.3390/biom14040409

  > Proteins need to be located in appropriate spatiotemporal contexts to carry out their diverse biological functions. Mislocalized proteins may lead to a broad range of diseases, such as cancer and Alzheimer's disease. Knowing where a target protein resides within a cell will give insights into tailored drug design for a disease. As the gold validation standard, the conventional wet lab uses fluorescent microscopy imaging, immunoelectron microscopy, and fluorescent biomarker tags for protein subcellular location identification. However, the booming era of proteomics and high-throughput sequencing generates tons of newly discovered proteins, making protein subcellular localization by wet-lab experiments a mission impossible. To tackle this concern, in the past decades, artificial intelligence (AI) and machine learning (ML), especially deep learning methods, have made significant progress in this research area. In this article, we review the latest advances in AI-based method development in three typical types of approaches, including sequence-based, knowledge-based, and image-based methods. We also elaborately discuss existing challenges and future directions in AI-based method development in this research field.

- Viana et al. (2023) — integrated intracellular organisation of human stem cells (Allen Institute).

  Integrated intracellular organization and its variations in human iPS cells. Viana et al. (2023). *Nature*. doi:10.1038/s41586-022-05563-7

  > Understanding how a subset of expressed genes dictates cellular phenotype is a considerable challenge owing to the large numbers of molecules involved, their combinatorics and the plethora of cellular behaviours that they determine 1,2 . Here we reduced this complexity by focusing on cellular organization-a key readout and driver of cell behaviour 3,4 -at the level of major cellular structures that represent distinct organelles and functional machines, and generated the WTC-11 hiPSC Single-Cell Image Dataset v1, which contains more than 200,000 live cells in 3D, spanning 25 key cellular structures. The scale and quality of this dataset permitted the creation of a generalizable analysis framework to convert raw image data of cells and their structures into dimensionally reduced, quantitative measurements that can be interpreted by humans, and to facilitate data exploration. This framework embraces the vast cell-to-cell variability that is observed within a normal population, facilitates the integration of cell-by-cell structural data and allows quantitative analyses of distinct, separable aspects of organization within and across different cell populations. We found that the integrated intracellular organization of interphase cells was robust to the wide range of variation in cell shape in the population; that the average locations of some structures became polarized in cells at the edges of colonies while maintaining the 'wiring' of their interactions with other structures; and that, by contrast, changes in the location of structures during early mitotic reorganization were accompanied by changes in their wiring.

B.13.6 Tissues and pathology

- Goltsev et al. (2018), Giesen et al. (2014), Angelo et al. (2014) — CODEX, imaging mass cytometry and MIBI.

Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Goltsev et al. (2018). *Cell*. doi:10.1016/j.cell.2018.07.010

> A highly multiplexed cytometric imaging approach, termed co-detection by indexing (CODEX), is used here to create multiplexed datasets of normal and lupus (MRL/lpr) murine spleens. CODEX iteratively visualizes antibody binding events using DNA barcodes, fluorescent dNTP analogs, and an in situ polymerization-based indexing procedure. An algorithmic pipeline for single-cell antigen quantification in tightly packed tissues was developed and used to overlay well-known morphological features with de novo characterization of lymphoid tissue architecture at a single-cell and cellular neighborhood levels. We observed an unexpected, profound impact of the cellular neighborhood on the expression of protein receptors on immune cells. By comparing normal murine spleen to spleens from animals with systemic autoimmune disease (MRL/lpr), extensive and previously uncharacterized splenic cell-interaction dynamics in the healthy versus diseased state was observed. The fidelity of multiplexed spatial cytometry demonstrated here allows for quantitative systemic characterization of tissue architecture in normal and clinically aberrant samples.

Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Giesen et al. (2014). *Nature Methods*. doi:10.1038/nmeth.2869

> Mass cytometry enables high-dimensional, single-cell analysis of cell type and state. In mass cytometry, rare earth metals are used as reporters on antibodies. Analysis of metal abundances using the mass cytometer allows determination of marker expression in individual cells. Mass cytometry has previously been applied only to cell suspensions. To gain spatial information, we have coupled immunohistochemical and immunocytochemical methods with high-resolution laser ablation to CyTOF mass cytometry. This approach enables the simultaneous imaging of 32 proteins and protein modifications at subcellular resolution; with the availability of additional isotopes, measurement of over 100 markers will be possible. We applied imaging mass cytometry to human breast cancer samples, allowing delineation of cell subpopulations and cell-cell interactions and highlighting tumor heterogeneity. Imaging mass cytometry complements existing imaging approaches. It will enable basic studies of tissue heterogeneity and function and support the transition of medicine toward individualized molecularly targeted diagnosis and therapies.

Multiplexed ion beam imaging of human breast tumors. Angelo et al. (2014). *Nature Medicine*. doi:10.1038/nm.3488

> Immunohistochemistry (IHC) is a tool for visualizing protein expression that is employed as part of the diagnostic workup for the majority of solid tissue malignancies. Existing IHC methods use antibodies tagged with fluorophores or enzyme reporters that generate colored pigments. Because these reporters exhibit spectral and spatial overlap when used simultaneously, multiplexed IHC is not routinely used in clinical settings. We have developed a method that uses secondary ion mass spectrometry to image antibodies tagged with isotopically pure elemental metal reporters. Multiplexed ion beam imaging (MIBI) is capable of analyzing up to 100 targets simultaneously over a five-log dynamic range. Here, we used MIBI to analyze formalin-fixed, paraffin-embedded human breast tumor tissue sections stained with ten labels simultaneously. The resulting data suggest that MIBI can provide new insights into disease pathogenesis that will be valuable for basic research, drug discovery and clinical diagnostics.

- Janesick et al. (2023), Park et al. (2022) — imaging-based spatial transcriptomics and an overview of spatial omics.

High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Janesick et al. (2023). *Nature Communications*. doi:10.1038/s41467-023-43458-x

> Single-cell and spatial technologies that profile gene expression across a whole tissue are revolutionizing the resolution of molecular states in clinical samples. Current commercially available technologies provide whole transcriptome single-cell, whole transcriptome spatial, or targeted in situ gene expression analysis. Here, we combine these technologies to explore tissue heterogeneity in large, FFPE human breast cancer sections. This integrative approach allowed us to explore molecular differences that exist between distinct tumor regions and to identify biomarkers involved in the progression towards invasive carcinoma. Further, we study cell neighborhoods and identify rare boundary cells that sit at the critical myoepithelial border confining the spread of malignant cells. Here, we demonstrate that each technology alone provides information about molecular signatures relevant to understanding cancer heterogeneity; however, it is the integration of these technologies that leads to deeper insights, ushering in discoveries that will progress oncology research and the development of diagnostics and therapeutics.

Spatial omics technologies at multimodal and single cell/subcellular level. Park et al. (2022). *Genome Biology*. doi:10.1186/s13059-022-02824-6

> Spatial omics technologies enable a deeper understanding of cellular organizations and interactions within a tissue of interest. These assays can identify specific compartments or regions in a tissue with differential transcript or protein abundance, delineate their interactions, and complement other methods in defining cellular phenotypes. A variety of spatial methodologies are being developed and commercialized; however, these techniques differ in spatial resolution, multiplexing capability, scale/throughput, and coverage. Here, we review the current and prospective landscape of single cell to subcellular resolution spatial omics technologies and analysis tools to provide a comprehensive picture for both research and clinical applications.

- Amitay et al. (2023) — classifying cells in multiplexed images.

CellSighter: a neural network to classify cells in highly multiplexed images. Amitay et al. (2023). *Nature Communications*. doi:10.1038/s41467-023-40066-7

> Multiplexed imaging enables measurement of multiple proteins in situ, offering an unprecedented opportunity to chart various cell types and states in tissues. However, cell classification, the task of identifying the type of individual cells, remains challenging, labor-intensive, and limiting to throughput. Here, we present CellSighter, a deep-learning based pipeline to accelerate cell classification in multiplexed images. Given a small training set of expert-labeled images, CellSighter outputs the label probabilities for all cells in new images. CellSighter achieves over 80% accuracy for major cell types across imaging platforms, which approaches inter-observer concordance. Ablation studies and simulations show that CellSighter is able to generalize its training data and learn features of protein expression levels, as well as spatial features such as subcellular expression patterns. CellSighter's design reduces overfitting, and it can be trained with only thousands or even hundreds of labeled examples. CellSighter also outputs a prediction confidence, allowing downstream experts control over the results. Altogether, CellSighter drastically reduces hands-on time for cell classification in multiplexed images, while improving accuracy and consistency across datasets.

- Chen et al. (2024), Lu et al. (2024), Vorontsov et al. (2024), Xu et al. (2024) — pathology foundation models.

Towards a general-purpose foundation model for computational pathology. Chen et al. (2024). *Nature Medicine*. doi:10.1038/s41591-024-02857-3

> Quantitative evaluation of tissue images is crucial for computational pathology (CPath) tasks, requiring the objective characterization of histopathological entities from whole-slide images (WSIs). The high resolution of WSIs and the variability of morphological features present significant challenges, complicating the large-scale annotation of data for high-performance applications. To address this challenge, current efforts have proposed the use of pretrained image encoders through transfer learning from natural image datasets or self-supervised learning on publicly available histopathology datasets, but have not been extensively developed and evaluated across diverse tissue types at scale. We introduce UNI, a general-purpose self-supervised model for pathology, pretrained using more than 100 million images from over 100,000 diagnostic H&E-stained WSIs (>77 TB of data) across 20 major tissue types. The model was evaluated on 34 representative CPath tasks of varying diagnostic difficulty. In addition to outperforming previous state-of-the-art models, we demonstrate new modeling capabilities in CPath such as resolution-agnostic tissue classification, slide classification using few-shot class prototypes, and disease subtyping generalization in classifying up to 108 cancer types in the OncoTree classification system. UNI advances unsupervised representation learning at scale in CPath in terms of both pretraining data and downstream evaluation, enabling data-efficient artificial intelligence models that can generalize and transfer to a wide range of diagnostically challenging tasks and clinical workflows in anatomic pathology.

A visual-language foundation model for computational pathology. Lu et al. (2024). *Nature Medicine*. doi:10.1038/s41591-024-02856-4

> The accelerated adoption of digital pathology and advances in deep learning have enabled the development of robust models for various pathology tasks across a diverse array of diseases and patient cohorts. However, model training is often difficult due to label scarcity in the medical domain, and a model's usage is limited by the specific task and disease for which it is trained. Additionally, most models in histopathology leverage only image data, a stark contrast to how humans teach each other and reason about histopathologic entities. We introduce CONtrastive learning from Captions for Histopathology

(CONCH), a visual-language foundation model developed using diverse sources of histopathology images, biomedical text and, notably, over 1.17 million image-caption pairs through task-agnostic pretraining. Evaluated on a suite of 14 diverse benchmarks, CONCH can be transferred to a wide range of downstream tasks involving histopathology images and/or text, achieving state-of-the-art performance on histology image classification, segmentation, captioning, and text-to-image and image-to-text retrieval. CONCH represents a substantial leap over concurrent visual-language pretrained systems for histopathology, with the potential to directly facilitate a wide array of machine learning-based workflows requiring minimal or no further supervised fine-tuning.

A foundation model for clinical-grade computational pathology and rare cancers detection. Vorontsov et al. (2024). *Nature Medicine*. doi:10.1038/s41591-024-03141-0

The analysis of histopathology images with artificial intelligence aims to enable clinical decision support systems and precision medicine. The success of such applications depends on the ability to model the diverse patterns observed in pathology images. To this end, we present Virchow, the largest foundation model for computational pathology to date. In addition to the evaluation of biomarker prediction and cell identification, we demonstrate that a large foundation model enables pan-cancer detection, achieving 0.95 specimen-level area under the (receiver operating characteristic) curve across nine common and seven rare cancers. Furthermore, we show that with less training data, the pan-cancer detector built on Virchow can achieve similar performance to tissue-specific clinical-grade models in production and outperform them on some rare variants of cancer. Virchow's performance gains highlight the value of a foundation model and open possibilities for many high-impact applications with limited amounts of labeled training data.

A whole-slide foundation model for digital pathology from real-world data. Xu et al. (2024). *Nature*. doi:10.1038/s41586-024-07441-w

Digital pathology poses unique computational challenges, as a standard gigapixel slide may comprise tens of thousands of image tiles 1-3 . Prior models have often resorted to subsampling a small portion of tiles for each slide, thus missing the important slide-level context 4 . Here we present Prov-GigaPath, a whole-slide pathology foundation model pretrained on 1.3 billion 256 × 256 pathology image tiles in 171,189 whole slides from Providence, a large US health network comprising 28 cancer centres. The slides originated from more than 30,000 patients covering 31 major tissue types. To pretrain Prov-GigaPath, we propose GigaPath, a novel vision transformer architecture for pretraining gigapixel pathology slides. To scale GigaPath for slide-level learning with tens of thousands of image tiles, GigaPath adapts the newly developed LongNet 5 method to digital pathology. To evaluate Prov-GigaPath, we construct a digital pathology benchmark comprising 9 cancer subtyping tasks and 17 pathomics tasks, using both Providence and TCGA data 6 . With large-scale pretraining and ultra-large-context modelling, Prov-GigaPath attains state-of-the-art performance on 25 out of 26 tasks, with significant improvement over the second-best method on 18 tasks. We further demonstrate the potential of Prov-GigaPath on vision-language pretraining for pathology 7,8 by incorporating the pathology reports. In sum, Prov-GigaPath is an open-weight foundation model that achieves state-of-the-art performance on various digital pathology tasks, demonstrating the importance of real-world data and whole-slide modelling.

- He et al. (2020) — predicting gene expression from histology.

Integrating spatial gene expression and breast tumour morphology via deep learning. He et al. (2020). *Nature Biomedical Engineering*. doi:10.1038/s41551-020-0578-x

Spatial transcriptomics allows for the measurement of RNA abundance at a high spatial resolution, making it possible to systematically link the morphology of cellular neighbourhoods and spatially localized gene expression. Here, we report the development of a deep learning algorithm for the prediction of local gene expression from haematoxylin-and-eosin-stained histopathology images using a new dataset of 30,612 spatially resolved gene expression data matched to histopathology images from 23 patients with breast cancer. We identified over 100 genes, including known breast cancer biomarkers of intratumoral heterogeneity and the co-localization of tumour growth and immune activation, the expression of which can be predicted from the histopathology images at a resolution of 100 μm. We also show that the algorithm generalizes well to The Cancer Genome Atlas and to other breast cancer gene expression datasets without the need for re-training. Predicting the spatially resolved transcriptome of a tissue directly from tissue images may enable image-based screening for molecular biomarkers with spatial variation.

- Howard et al. (2021) — site signatures and bias in pathology models.

  The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Howard et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-24698-1

  > The Cancer Genome Atlas (TCGA) is one of the largest biorepositories of digital histology. Deep learning (DL) models have been trained on TCGA to predict numerous features directly from histology, including survival, gene expression patterns, and driver mutations. However, we demonstrate that these features vary substantially across tissue submitting sites in TCGA for over 3,000 patients with six cancer subtypes. Additionally, we show that histologic image differences between submitting sites can easily be identified with DL. Site detection remains possible despite commonly used color normalization and augmentation methods, and we quantify the image characteristics constituting this site-specific digital histology signature. We demonstrate that these site-specific signatures lead to biased accuracy for prediction of features including survival, genomic mutations, and tumor stage. Furthermore, ethnicity can also be inferred from site-specific signatures, which must be accounted for to ensure equitable application of DL. These site-specific signatures can lead to overoptimistic estimates of model performance, and we propose a quadratic programming method that abrogates this bias by ensuring models are not trained and validated on samples from the same site.

B.13.7 Atlases and virtual cells

- Regev et al. (2017), Rood et al. (2024) — the Human Cell Atlas.

  The Human Cell Atlas. Regev et al. (2017). *eLife*. doi:10.7554/elife.27041

  > The recent advent of methods for high-throughput single-cell molecular profiling has catalyzed a growing sense in the scientific community that the time is ripe to complete the 150-year-old effort to identify all cell types in the human body. The Human Cell Atlas Project is an international collaborative effort that aims to define all human cell types in terms of distinctive molecular profiles (such as gene expression profiles) and to connect this information with classical cellular descriptions (such as location and morphology). An open comprehensive reference map of the molecular state of cells in healthy human tissues would propel the systematic study of physiological states, developmental trajectories, regulatory circuitry and interactions of cells, and also provide a framework for understanding cellular dysregulation in human disease. Here we describe the idea, its potential utility, early proofs-of-concept, and some design considerations for the Human Cell Atlas, including a commitment to open data, code, and community.

  The Human Cell Atlas from a cell census to a unified foundation model. Rood et al. (2024). *Nature*. doi:10.1038/s41586-024-08338-4

  > With the convergence of notable advances in molecular and spatial profiling methods and new computational approaches taking advantage of artificial intelligence and machine learning, the construction of cell atlases is progressing from data collection to atlas integration and beyond. Here, we explore five ways in which cell atlases, including the Human Cell Atlas, are already revealing valuable biological insights, and how they are poised to provide even greater benefits in the coming years. In particular, we discuss cell atlases as censuses of cells; as 3D maps of cells in the body, across modalities and scales; as maps connecting genotype causes to phenotype effects; as 4D maps of development; and, ultimately, as foundation models of biology unifying all of these aspects and helping to transform medicine.

- HuBMAP Consortium (2019), Jain et al. (2023) — HuBMAP.

  The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. HuBMAP Consortium (2019). *Nature*. doi:10.1038/s41586-019-1629-x

  > Transformative technologies are enabling the construction of three-dimensional maps of tissues with unprecedented spatial and molecular resolution. Over the next seven years, the NIH Common Fund Human Biomolecular Atlas Program (HuBMAP) intends to develop a widely accessible framework for comprehensively mapping the human body at single-cell resolution by supporting technology development, data acquisition, and detailed spatial mapping. HuBMAP will integrate its efforts with other funding agencies, programs, consortia, and the biomedical research community at large towards the shared vision of a comprehensive, accessible three-dimensional molecular and cellular atlas of the human body, in health and under various disease conditions.

Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). Jain et al. (2023). *Nature Cell Biology*. doi:10.1038/s41556-023-01194-w

> The Human BioMolecular Atlas Program (HuBMAP) aims to create a multi-scale spatial atlas of the healthy human body at single-cell resolution by applying advanced technologies and disseminating resources to the community. As the HuBMAP moves past its first phase, creating ontologies, protocols and pipelines, this Perspective introduces the production phase: the generation of reference spatial maps of functional tissue units across many organs from diverse populations and the creation of mapping tools and infrastructure to advance biomedical research.

- Pearce et al. (2025), Istrate et al. (2025) — TranscriptFormer and rBio (CZI; preprints).

A Cross-Species Generative Cell Atlas Across 1.5 Billion Years of Evolution: The TranscriptFormer Single-cell Model. Pearce et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.04.25.650731

> Single-cell transcriptomics has revolutionized our understanding of cellular diversity, yet our understanding of the transcriptional programs across the tree of life remains limited. Here we present TranscriptFormer, a family of generative foundation models trained on up to 112 million cells spanning 1.53 billion years of evolution across 12 species. By jointly modeling gene identities and expression levels using a novel generative architecture, TranscriptFormer encodes multi-scale biological structure, functioning as a queryable virtual cell atlas. We demonstrate state-of-the-art performance on both in-distribution and out-of-distribution cell type classification, with robust performance even for species separated by over 685 million years of evolution. TranscriptFormer can also perform zero-shot disease state identification in human cells and accurately transfers cell state annotations across species boundaries. As a generative model, TranscriptFormer can be prompted to predict cell type-specific transcription factors and gene-gene interactions that align with independent experimental observations. Developmental trajectories, phylogenetic relationships and cellular hierarchies emerge naturally in TranscriptFormer's representations without any explicit training on these annotations. This work establishes a powerful framework for quantitative single-cell analysis, and comparative cellular biology, thus demonstrating that universal principles of cellular organization can be learned and predicted across the tree of life.

rbio1 - training scientific reasoning LLMs with biological world models as soft verifiers. Istrate et al. (2025). *bioRxiv (preprint)*. doi:10.1101/2025.08.18.670981

> Reasoning models are typically trained against verification mechanisms in formally specified systems such as code or symbolic math. In open domains like biology, however, we lack exact rules to enable large-scale formal verification and instead often rely on lab experiments to test predictions. Such experiments are slow, costly, and cannot scale with computation. In this work, we show that world models of biology or other prior knowledge can serve as approximate oracles for soft verification , allowing reasoning systems to be trained without additional experimental data. We present two paradigms of training models with approximate verifiers: RLEMF : reinforcement learning with experimental model feedback and RLPK : reinforcement learning from prior knowledge. Using these paradigms, we introduce rbio1 , a reasoning model for biology post-trained from a pretrained LLM with reinforcement learning, using learned biological models for verification during training. We demonstrate that soft verification can distill biological world models into rbio1 , enabling it to achieve state-of-the-art performance on perturbation prediction in the PerturbQA benchmark. We further show that composing multiple AI-verifiers improves performance and that models trained with soft biological rewards transfer zero-shot to cross-domain tasks such as disease-state prediction. We present rbio1 as a proof of concept that predictions from biological models can train powerful reasoning systems using simulations rather than experimental data, offering a new paradigm for model training.

- Roohani et al. (2023), Adduri et al. (2026), Roohani et al. (2025) — GEARS, State and the Virtual Cell Challenge.

Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Roohani et al. (2023). *Nature Biotechnology*. doi:10.1038/s41587-023-01905-6

> Understanding cellular responses to genetic perturbation is central to numerous biomedical applications, from identifying genetic interactions involved in cancer to developing methods for regenerative medicine. However, the combinatorial explosion in the number of possible multigene perturbations severely limits experimental interrogation. Here, we present graph-enhanced gene activation and repression simulator (GEARS), a method that integrates deep learning with a knowledge graph of gene-gene

relationships to predict transcriptional responses to both single and multigene perturbations using single-cell RNA-sequencing data from perturbational screens. GEARS is able to predict outcomes of perturbing combinations consisting of genes that were never experimentally perturbed. GEARS exhibited 40% higher precision than existing approaches in predicting four distinct genetic interaction subtypes in a combinatorial perturbation screen and identified the strongest interactions twice as well as prior approaches. Overall, GEARS can predict phenotypically distinct effects of multigene perturbations and thus guide the design of perturbational experiments.

Predicting cellular responses to perturbation across diverse contexts with State. Adduri et al. (2026). *Cell.* doi:10.1016/j.cell.2026.07.052

> While machine learning models offer potential for predicting transcriptomic effects of perturbation, they currently struggle to generalize across cellular contexts. Here, we introduce State, a machine learning model that predicts perturbation effects while accounting for cellular heterogeneity within and across experiments. State is trained using single-cell gene expression data to predict perturbation effects across sets of cells. State improved discrimination of effects on large datasets by more than 30% and identified differentially expressed genes across genetic, signaling, and chemical perturbations with significantly improved accuracy compared with baselines. Its cell embeddings trained on observational data from 167 million cells enable the identification of strong perturbations in cellular contexts where no perturbations were observed during training. We further introduce Cell-Eval, a comprehensive evaluation framework that can be used to evaluate future models. Overall, the performance and flexibility of State set the stage for scaling the development of AI models of cell state.

Virtual Cell Challenge: Toward a Turing test for the virtual cell. Roohani et al. (2025). *Cell.* doi:10.1016/j.cell.2025.06.008

> Virtual cells are an emerging frontier at the intersection of artificial intelligence and biology. A key goal of these cell state models is predicting cellular responses to perturbations. The Virtual Cell Challenge is being established to catalyze progress toward this goal. This recurring and open benchmark competition from the Arc Institute will provide an evaluation framework, purpose-built datasets, and a venue for accelerating model development.

- Kedzierska et al. (2025) — zero-shot evaluation of single-cell foundation models.

Zero-shot evaluation reveals limitations of single-cell foundation models. Kedzierska et al. (2025). *Genome Biology.* doi:10.1186/s13059-025-03574-x

> Foundation models such as scGPT and Geneformer have not been rigorously evaluated in a setting where they are used without any further training (i.e., zero-shot). Understanding the performance of models in zero-shot settings is critical to applications that exclude the ability to fine-tune, such as discovery settings where labels are unknown. Our evaluation of the zero-shot performance of Geneformer and scGPT suggests that, in some cases, these models may face reliability challenges and could be outperformed by simpler methods. Our findings underscore the importance of zero-shot evaluations in development and deployment of foundation models in single-cell research.

B.13.8 Whole-cell models and cryo-electron tomography

- Karr et al. (2012) — the first whole-cell model.

A Whole-Cell Computational Model Predicts Phenotype from Genotype. Karr et al. (2012). *Cell.* doi:10.1016/j.cell.2012.05.044

> Understanding how complex phenotypes arise from individual molecules and their interactions is a primary challenge in biology that computational approaches are poised to tackle. We report a whole-cell computational model of the life cycle of the human pathogen Mycoplasma genitalium that includes all of its molecular components and their interactions. An integrative approach to modeling that combines diverse mathematics enabled the simultaneous inclusion of fundamentally different cellular processes and experimental measurements. Our whole-cell model accounts for all annotated gene functions and was validated against a broad range of data. The model provides insights into many previously unobserved cellular behaviors, including in vivo rates of protein-DNA association and an inverse relationship between the durations of DNA replication initiation and replication. In addition, experimental analysis directed

by model predictions identified previously undetected kinetic parameters and biological functions. We conclude that comprehensive whole-cell models can be used to facilitate biological discovery.

- Thornburg et al. (2022) — a kinetic model of a living minimal cell.

  Fundamental behaviors emerge from simulations of a living minimal cell. Thornburg et al. (2022). *Cell*. doi:10.1016/j.cell.2021.12.025

  > We present a whole-cell fully dynamical kinetic model (WCM) of JCVI-syn3A, a minimal cell with a reduced genome of 493 genes that has retained few regulatory proteins or small RNAs. Cryo-electron tomograms provide the cell geometry and ribosome distributions. Time-dependent behaviors of concentrations and reaction fluxes from stochastic-deterministic simulations over a cell cycle reveal how the cell balances demands of its metabolism, genetic information processes, and growth, and offer insight into the principles of life for this minimal cell. The energy economy of each process including active transport of amino acids, nucleosides, and ions is analyzed. WCM reveals how emergent imbalances lead to slowdowns in the rates of transcription and translation. Integration of experimental data is critical in building a kinetic model from which emerges a genome-wide distribution of mRNA half-lives, multiple DNA replication events that can be compared to qPCR results, and the experimentally observed doubling behavior.

- Maritan et al. (2022), Stevens et al. (2023), Yu et al. (2016) — structural models and molecular dynamics of cells.

  Building Structural Models of a Whole Mycoplasma Cell. Maritan et al. (2022). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2021.167351

  > Building structural models of entire cells has been a long-standing cross-discipline challenge for the research community, as it requires an unprecedented level of integration between multiple sources of biological data and enhanced methods for computational modeling and visualization. Here, we present the first 3D structural models of an entire Mycoplasma genitalium (MG) cell, built using the CellPACK suite of computational modeling tools. Our model recapitulates the data described in recent whole-cell system biology simulations and provides a structural representation for all MG proteins, DNA and RNA molecules, obtained by combining experimental and homology-modeled structures and lattice-based models of the genome. We establish a framework for gathering, curating and evaluating these structures, exposing current weaknesses of modeling methods and the boundaries of MG structural knowledge, and visualization methods to explore functional characteristics of the genome and proteome. We compare two approaches for data gathering, a manually-curated workflow and an automated workflow that uses homologous structures, both of which are appropriate for the analysis of mesoscale properties such as crowding and volume occupancy. Analysis of model quality provides estimates of the regularization that will be required when these models are used as starting points for atomic molecular dynamics simulations.

  Molecular dynamics simulation of an entire cell. Stevens et al. (2023). *Frontiers in Chemistry*. doi:10.3389/fchem.2023.1106495

  > The ultimate microscope, directed at a cell, would reveal the dynamics of all the cell's components with atomic resolution. In contrast to their real-world counterparts, computational microscopes are currently on the brink of meeting this challenge. In this perspective, we show how an integrative approach can be employed to model an entire cell, the minimal cell, JCVI-syn3A, at full complexity. This step opens the way to interrogate the cell's spatio-temporal evolution with molecular dynamics simulations, an approach that can be extended to other cell types in the near future.

  Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. Yu et al. (2016). *eLife*. doi:10.7554/elife.19274

  > Biological macromolecules function in highly crowded cellular environments. The structure and dynamics of proteins and nucleic acids are well characterized in vitro, but in vivo crowding effects remain unclear. Using molecular dynamics simulations of a comprehensive atomistic model cytoplasm we found that protein-protein interactions may destabilize native protein structures, whereas metabolite interactions may induce more compact states due to electrostatic screening. Protein-protein interactions also resulted in significant variations in reduced macromolecular diffusion under crowded conditions, while metabolites exhibited significant two-dimensional surface diffusion and altered protein-ligand binding that may reduce the effective concentration of metabolites and ligands in vivo. Metabolic enzymes showed weak non-specific association in cellular environments attributed to solvation and entropic effects. These effects are

> expected to have broad implications for the in vivo functioning of biomolecules. This work is a first step towards physically realistic in silico whole-cell models that connect molecular with cellular biology.

- Nickell et al. (2006), Beck and Baumeister (2016), Young and Villa (2023) — visual proteomics and cryo-electron tomography.

  A visual approach to proteomics. Nickell et al. (2006). *Nature Reviews Molecular Cell Biology*. doi:10.1038/nrm1861

  > Cryo-electron tomography is an emerging imaging technique that has unique potential for molecular cell biology. At the present resolution of 4-5 nm, large supramolecular structures can be studied in unperturbed cellular environments and, in the future, it will become possible to map molecular landscapes inside cells in a more comprehensive manner. 'Visual proteomics' aims to complement and extend mass-spectrometry-based inventories, and to provide a quantitative description of the macromolecular interactions that underlie cellular functions.

  Cryo-Electron Tomography: Can it Reveal the Molecular Sociology of Cells in Atomic Detail?. Beck & Baumeister (2016). *Trends in Cell Biology*. doi:10.1016/j.tcb.2016.08.006

  > Traditionally, macromolecular structure determination is performed ex situ, that is, with purified materials. But, there are strong incentives to develop approaches to study them in situ in their native functional context. In recent years, cryo-electron tomography (cryo-ET) has emerged as a powerful method for visualizing the molecular organization of unperturbed cellular landscapes with the potential to attain near-atomic resolution. Here, we review recent work on several macromolecular assemblies, demonstrating the power of in situ studies. We also highlight technical challenges and discuss ways to meet them.

  Bringing Structure to Cell Biology with Cryo-Electron Tomography. Young & Villa (2023). *Annual Review of Biophysics*. doi:10.1146/annurev-biophys-111622-091327

  > Recent advances in cryo-electron microscopy have marked only the beginning of the potential of this technique. To bring structure into cell biology, the modality of cryo-electron tomography has fast developed into a bona fide in situ structural biology technique where structures are determined in their native environment, the cell. Nearly every step of the cryo-focused ion beam-assisted electron tomography (cryo-FIB-ET) workflow has been improved upon in the past decade, since the first windows were carved into cells, unveiling macromolecular networks in near-native conditions. By bridging structural and cell biology, cryo-FIB-ET is advancing our understanding of structure-function relationships in their native environment and becoming a tool for discovering new biology.

- O'Reilly et al. (2020), Tegunov et al. (2021) — structural biology inside *Mycoplasma pneumoniae* cells.

  In-cell architecture of an actively transcribing-translating expressome. O'Reilly et al. (2020). *Science*. doi:10.1126/science.abb3758

  > Structural biology studies performed inside cells can capture molecular machines in action within their native context. In this work, we developed an integrative in-cell structural approach using the genome-reduced human pathogen Mycoplasma pneumoniae We combined whole-cell cross-linking mass spectrometry, cellular cryo-electron tomography, and integrative modeling to determine an in-cell architecture of a transcribing and translating expressome at subnanometer resolution. The expressome comprises RNA polymerase (RNAP), the ribosome, and the transcription elongation factors NusG and NusA. We pinpointed NusA at the interface between a NusG-bound elongating RNAP and the ribosome and propose that it can mediate transcription-translation coupling. Translation inhibition dissociated the expressome, whereas transcription inhibition stalled and rearranged it. Thus, the active expressome architecture requires both translation and transcription elongation within the cell.

  Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. Tegunov et al. (2021). *Nature Methods*. doi:10.1038/s41592-020-01054-7

> Cryo-electron microscopy (cryo-EM) enables macromolecular structure determination in vitro and inside cells. In addition to aligning individual particles, accurate registration of sample motion and three-dimensional deformation during exposures are crucial for achieving high-resolution reconstructions. Here we describe M, a software tool that establishes a reference-based, multi-particle refinement framework for cryo-EM data and couples a comprehensive spatial deformation model to in silico correction of electron-optical aberrations. M provides a unified optimization framework for both frame-series and tomographic tilt-series data. We show that tilt-series data can provide the same resolution as frame-series data on a purified protein specimen, indicating that the alignment step no longer limits the resolution obtainable from tomographic data. In combination with Warp and RELION, M resolves to residue level a 70S ribosome bound to an antibiotic inside intact bacterial cells. Our work provides a computational tool that facilitates structural biology in cells.

- Teresa-Trueba et al. (2023), Rice et al. (2023), Lamm et al. (2026) — deep learning for finding particles and membranes in tomograms.

Convolutional networks for supervised mining of molecular patterns within cellular context. de Teresa-Trueba et al. (2023). *Nature Methods*. doi:10.1038/s41592-022-01746-2

> Cryo-electron tomograms capture a wealth of structural information on the molecular constituents of cells and tissues. We present DeePiCt (deep picker in context), an open-source deep-learning framework for supervised segmentation and macromolecular complex localization in cryo-electron tomography. To train and benchmark DeePiCt on experimental data, we comprehensively annotated 20 tomograms of Schizosaccharomyces pombe for ribosomes, fatty acid synthases, membranes, nuclear pore complexes, organelles, and cytosol. By comparing DeePiCt to state-of-the-art approaches on this dataset, we show its unique ability to identify low-abundance and low-density complexes. We use DeePiCt to study compositionally distinct subpopulations of cellular ribosomes, with emphasis on their contextual association with mitochondria and the endoplasmic reticulum. Finally, applying pre-trained networks to a HeLa cell tomogram demonstrates that DeePiCt achieves high-quality predictions in unseen datasets from different biological species in a matter of minutes. The comprehensively annotated experimental data and pre-trained networks are provided for immediate use by the community.

TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining. Rice et al. (2023). *Nature Methods*. doi:10.1038/s41592-023-01878-z

> Cryogenic-electron tomography enables the visualization of cellular environments in extreme detail, however, tools to analyze the full amount of information contained within these densely packed volumes are still needed. Detailed analysis of macromolecules through subtomogram averaging requires particles to first be localized within the tomogram volume, a task complicated by several factors including a low signal to noise ratio and crowding of the cellular space. Available methods for this task suffer either from being error prone or requiring manual annotation of training data. To assist in this crucial particle picking step, we present TomoTwin: an open source general picking model for cryogenic-electron tomograms based on deep metric learning. By embedding tomograms in an information-rich, high-dimensional space that separates macromolecules according to their three-dimensional structure, TomoTwin allows users to identify proteins in tomograms de novo without manually creating training data or retraining the network to locate new proteins.

MemBrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography. Lamm et al. (2026). *Nature Methods*. doi:10.1038/s41592-026-03178-8

> Cryo-electron tomography provides unique insights into macromolecular complexes in their native environments, yet membrane analysis remains a major bottleneck due to low signal-to-noise ratios, missing wedge artifacts and the complexity of membrane-associated particles. Existing tools often require extensive manual annotation, struggle with generalization across datasets and lack integrated solutions for segmentation, particle localization and quantitative analysis. We introduce MemBrain v2, a deep-learning-enabled framework that unifies these tasks into a streamlined pipeline. MemBrain-seg leverages a diverse, collaboratively generated training dataset and specialized model training strategies to achieve generalizable membrane segmentation across variable tomographic conditions. MemBrain-pick enables data-efficient localization of membrane-bound particles by integrating geometric constraints with deep learning, reducing the need for extensive manual annotation. MemBrain-stats provides quantitative insights into particle distributions, computing spatial metrics to analyze intramembrane particle organization. MemBrain v2 integrates seamlessly into cryo-electron tomography workflows, providing an accessible and structured approach to membrane analysis.

- Ermel et al. (2024), Peck et al. (2026) — the CryoET Data Portal and a particle-picking challenge.

  A data portal for providing standardized annotations for cryo-electron tomography. Ermel et al. (2024). *Nature Methods*. doi:10.1038/s41592-024-02477-2

  Lessons learned from a Kaggle challenge for particle picking in cryo-electron tomography. Peck et al. (2026). *Nature Methods*. doi:10.1038/s41592-026-03198-4

  > The difficulty of particle picking in cryo-electron tomography remains a barrier to routine in situ structure determination. Machine learning is well suited to overcome this bottleneck with efficient algorithms that generalize across molecular species. To spur new algorithm development, we held a 3-month Kaggle challenge that tasked contestants with annotating five molecular species across hundreds of experimental tomograms. Here we analyze the results of this competition, which successfully engaged >1,000 participants and delivered particle pickers that outperformed existing state of the art. Systematic comparisons of the contestants' submissions revealed the tolerance of subtomogram averaging to moderate but not severe over-picking and underscored the need for more robust measures of annotation quality. The winning models also highlighted the importance of data augmentation to overcome limited training data. All competition tomograms along with the ground truth and winning teams' annotations have been released on the CryoET Data Portal as a resource to benchmark current and future particle picking algorithms.

B.13.9 Tools, data and standards

- Schindelin et al. (2012), Rueden et al. (2017) — Fiji and ImageJ2.

  Fiji: an open-source platform for biological-image analysis. Schindelin et al. (2012). *Nature Methods*. doi:10.1038/nmeth.2019

  > Fiji is a distribution of the popular open-source software ImageJ focused on biological-image analysis. Fiji uses modern software engineering practices to combine powerful software libraries with a broad range of scripting languages to enable rapid prototyping of image-processing algorithms. Fiji facilitates the transformation of new algorithms into ImageJ plugins that can be shared with end users through an integrated update system. We propose Fiji as a platform for productive collaboration between computer science and biology research communities.

  ImageJ2: ImageJ for the next generation of scientific image data. Rueden et al. (2017). *BMC Bioinformatics*. doi:10.1186/s12859-017-1934-z

  > Background ImageJ is an image analysis program extensively used in the biological sciences and beyond. Due to its ease of use, recordable macro language, and extensible plug-in architecture, ImageJ enjoys contributions from non-programmers, amateur programmers, and professional developers alike. Enabling such a diversity of contributors has resulted in a large community that spans the biological and physical sciences. However, a rapidly growing user base, diverging plugin suites, and technical limitations have revealed a clear need for a concerted software engineering effort to support emerging imaging paradigms, to ensure the software's ability to handle the requirements of modern science. Results We rewrote the entire ImageJ codebase, engineering a redesigned plugin mechanism intended to facilitate extensibility at every level, with the goal of creating a more powerful tool that continues to serve the existing community while addressing a wider range of scientific requirements. This next-generation ImageJ, called "ImageJ2" in places where the distinction matters, provides a host of new functionality. It separates concerns, fully decoupling the data model from the user interface. It emphasizes integration with external applications to maximize interoperability. Its robust new plugin framework allows everything from image formats, to scripting languages, to visualization to be extended by the community. The redesigned data model supports arbitrarily large, N-dimensional datasets, which are increasingly common in modern image acquisition. Despite the scope of these changes, backwards compatibility is maintained such that this new functionality can be seamlessly integrated with the classic ImageJ interface, allowing users and developers to migrate to these new methods at their own pace. Conclusions Scientific imaging benefits from open-source programs that advance new method development and deployment to a diverse audience. ImageJ has continuously evolved with this idea in mind; however, new and emerging scientific requirements have posed corresponding challenges for ImageJ's development. The described improvements provide a framework engineered for flexibility, intended to support these requirements as well as accommodate future needs. Future efforts will focus on implementing new algorithms in this framework and expanding collaborations with other popular scientific software suites.

- Carpenter et al. (2006), Stirling et al. (2021) — CellProfiler.

  CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Carpenter et al. (2006). *Genome Biology*. doi:10.1186/gb-2006-7-10-r100

  > Biologists can now prepare and image thousands of samples per day using automation, enabling chemical screens and functional genomics (for example, using RNA interference). Here we describe the first free, open-source system designed for flexible, high-throughput cell image analysis, CellProfiler. CellProfiler can address a variety of biological questions quantitatively, including standard assays (for example, cell count, size, per-cell protein levels) and complex morphological assays (for example, cell/organelle shape or subcellular patterns of DNA or protein staining).

  CellProfiler 4: improvements in speed, utility and usability. Stirling et al. (2021). *BMC Bioinformatics*. doi:10.1186/s12859-021-04344-9

  > Background Imaging data contains a substantial amount of information which can be difficult to evaluate by eye. With the expansion of high throughput microscopy methodologies producing increasingly large datasets, automated and objective analysis of the resulting images is essential to effectively extract biological information from this data. CellProfiler is a free, open source image analysis program which enables researchers to generate modular pipelines with which to process microscopy images into interpretable measurements. Results Herein we describe CellProfiler 4, a new version of this software with expanded functionality. Based on user feedback, we have made several user interface refinements to improve the usability of the software. We introduced new modules to expand the capabilities of the software. We also evaluated performance and made targeted optimizations to reduce the time and cost associated with running common large-scale analysis pipelines. Conclusions CellProfiler 4 provides significantly improved performance in complex workflows compared to previous versions. This release will ensure that researchers will have continued access to CellProfiler's powerful computational tools in the coming years.

- Berg et al. (2019), Bankhead et al. (2017) — ilastik and QuPath.

  ilastik: interactive machine learning for (bio)image analysis. Berg et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0582-9

  > We present ilastik, an easy-to-use interactive tool that brings machine-learning-based (bio)image analysis to end users without substantial computational expertise. It contains pre-defined workflows for image segmentation, object classification, counting and tracking. Users adapt the workflows to the problem at hand by interactively providing sparse training annotations for a nonlinear classifier. ilastik can process data in up to five dimensions (3D, time and number of channels). Its computational back end runs operations on-demand wherever possible, allowing for interactive prediction on data larger than RAM. Once the classifiers are trained, ilastik workflows can be applied to new data from the command line without further user interaction. We describe all ilastik workflows in detail, including three case studies and a discussion on the expected performance.

  QuPath: Open source software for digital pathology image analysis. Bankhead et al. (2017). *Scientific Reports*. doi:10.1038/s41598-017-17204-5

  > QuPath is new bioimage analysis software designed to meet the growing need for a user-friendly, extensible, open-source solution for digital pathology and whole slide image analysis. In addition to offering a comprehensive panel of tumor identification and high-throughput biomarker evaluation tools, QuPath provides researchers with powerful batch-processing and scripting functionality, and an extensible platform with which to develop and share new algorithms to analyze complex tissue images. Furthermore, QuPath's flexible design makes it suitable for a wide range of additional image analysis applications across biomedical research.

- Chamier et al. (2021), Ouyang et al. (2022) — deep learning without programming (ZeroCostDL4Mic) and the BioImage Model Zoo (preprint).

  Democratising deep learning for microscopy with ZeroCostDL4Mic. von Chamier et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-22518-0

Deep Learning (DL) methods are powerful analytical tools for microscopy and can outperform conventional image processing pipelines. Despite the enthusiasm and innovations fuelled by DL technology, the need to access powerful and compatible resources to train DL networks leads to an accessibility barrier that novice users often find difficult to overcome. Here, we present ZeroCostDL4Mic, an entry-level platform simplifying DL access by leveraging the free, cloud-based computational resources of Google Colab. ZeroCostDL4Mic allows researchers with no coding expertise to train and apply key DL networks to perform tasks including segmentation (using U-Net and StarDist), object detection (using YOLOv2), denoising (using CARE and Noise2Void), super-resolution microscopy (using Deep-STORM), and image-to-image translation (using Label-free prediction - fnet, pix2pix and CycleGAN). Importantly, we provide suitable quantitative tools for each network to evaluate model performance, allowing model optimisation. We demonstrate the application of the platform to study multiple biological processes.

BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis. Ouyang et al. (2022). *bioRxiv (preprint)*. doi:10.1101/2022.06.07.495102

Deep learning-based approaches are revolutionizing imaging-driven scientific research. However, the accessibility and reproducibility of deep learning-based workflows for imaging scientists remain far from sufficient. Several tools have recently risen to the challenge of democratizing deep learning by providing user-friendly interfaces to analyze new data with pre-trained or fine-tuned models. Still, few of the existing pre-trained models are interoperable between these tools, critically restricting a model's overall utility and the possibility of validating and reproducing scientific analyses. Here, we present the BioImage Model Zoo (https://bioimage.io): a community-driven, fully open resource where standardized pre-trained models can be shared, explored, tested, and downloaded for further adaptation or direct deployment in multiple end user-facing tools (e.g., ilastik, deepImageJ, QuPath, StarDist, ImJoy, ZeroCostDL4Mic, CSBDeep). To enable everyone to contribute and consume the Zoo resources, we provide a model standard to enable cross-compatibility, a rich list of example models and practical use-cases, developer tools, documentation, and the accompanying infrastructure for model upload, download and testing. Our contribution aims to lay the groundwork to make deep learning methods for microscopy imaging findable, accessible, interoperable, and reusable (FAIR) across software tools and platforms.

- Ljosa et al. (2012), Williams et al. (2017), Hartley et al. (2022) — BBBC, the Image Data Resource and the BioImage Archive.

  Annotated high-throughput microscopy image sets for validation. Ljosa et al. (2012). *Nature Methods*. doi:10.1038/nmeth.2083

  Image Data Resource: a bioimage data integration and publication platform. Williams et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4326

  Access to primary research data is vital for the advancement of science. To extend the data types supported by community repositories, we built a prototype Image Data Resource (IDR) that collects and integrates imaging data acquired across many different imaging modalities. IDR links data from several imaging modalities, including high-content screening, super-resolution and time-lapse microscopy, digital pathology, public genetic or chemical databases, and cell and tissue phenotypes expressed using controlled ontologies. Using this integration, IDR facilitates the analysis of gene networks and reveals functional interactions that are inaccessible to individual studies. To enable re-analysis, we also established a computational resource based on Jupyter notebooks that allows remote access to the entire IDR. IDR is also an open source platform that others can use to publish their own image data. Thus IDR provides both a novel on-line resource and a software infrastructure that promotes and extends publication and re-analysis of scientific image data.

  The BioImage Archive – Building a Home for Life-Sciences Microscopy Data. Hartley et al. (2022). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2022.167505

  Despite the huge impact of data resources in genomics and structural biology, until now there has been no central archive for biological data for all imaging modalities. The BioImage Archive is a new data resource at the European Bioinformatics Institute (EMBL-EBI) designed to fill this gap. In its initial development BioImage Archive accepts bioimaging data associated with publications, in any format, from any imaging modality from the molecular to the organism scale, excluding medical imaging. The BioImage Archive will ensure reproducibility of published studies that derive results from image data and reduce duplication of effort. Most importantly, the BioImage Archive will help scientists to generate new insights through reuse of existing data to answer new biological questions, and provision of training,

testing and benchmarking data for development of tools for image analysis. The archive is available at https://www.ebi.ac.uk/bioimage-archive/.

- Sarkans et al. (2021), Moore et al. (2021), Moore et al. (2023) — REMBI metadata and the OME-Zarr format.

  REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. Sarkans et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01166-8

  OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. Moore et al. (2021). *Nature Methods*. doi:10.1038/s41592-021-01326-w

  > The rapid pace of innovation in biological imaging and the diversity of its applications have prevented the establishment of a community-agreed standardized data format. We propose that complementing established open formats such as OME-TIFF and HDF5 with a next-generation file format such as Zarr will satisfy the majority of use cases in bioimaging. Critically, a common metadata format used in all these vessels can deliver truly findable, accessible, interoperable and reusable bioimaging data.

  OME-Zarr: a cloud-optimized bioimaging file format with international community support. Moore et al. (2023). *Histochemistry and Cell Biology*. doi:10.1007/s00418-023-02209-1

  > A growing community is constructing a next-generation file format (NGFF) for bioimaging to overcome problems of scalability and heterogeneity. Organized by the Open Microscopy Environment (OME), individuals and institutes across diverse modalities facing these problems have designed a format specification process (OME-NGFF) to address these needs. This paper brings together a wide range of those community members to describe the cloud-optimized format itself-OME-Zarr-along with tools and data resources available today to increase FAIR access and remove barriers in the scientific process. The current momentum offers an opportunity to unify a key component of the bioimaging domain-the file format that underlies so many personal, institutional, and global data management and analysis tasks.

- Maška et al. (2023) — the Cell Tracking Challenge.

  The Cell Tracking Challenge: 10 years of objective benchmarking. Maška et al. (2023). *Nature Methods*. doi:10.1038/s41592-023-01879-y

  > The Cell Tracking Challenge is an ongoing benchmarking initiative that has become a reference in cell segmentation and tracking algorithm development. Here, we present a significant number of improvements introduced in the challenge since our 2017 report. These include the creation of a new segmentation-only benchmark, the enrichment of the dataset repository with new datasets that increase its diversity and complexity, and the creation of a silver standard reference corpus based on the most competitive results, which will be of particular interest for data-hungry deep learning-based strategies. Furthermore, we present the up-to-date cell segmentation and tracking leaderboards, an in-depth analysis of the relationship between the performance of the state-of-the-art methods and the properties of the datasets and annotations, and two novel, insightful studies about the generalizability and the reusability of top-performing methods. These studies provide critical practical conclusions for both developers and users of traditional and machine learning-based cell segmentation and tracking algorithms.

C Textbooks and general resources

C.1 Textbooks

No single textbook is required. These cover most of the material, but none is fully up to date.

- Zvelebil and Baum (2008) — *Understanding Bioinformatics.* Broad coverage, recommended in the exam description.
- Pevsner (2015) — *Bioinformatics and Functional Genomics.* Broad coverage; the basis of the 2021 exam variant.
- Jones and Pevzner (2004) — *An Introduction to Bioinformatics Algorithms.* Algorithms: alignment, suffix trees, assembly, phylogeny, motifs.
- Compeau and Pevzner (2018) — *Bioinformatics Algorithms: An Active Learning Approach.* With free online lectures.
- Durbin et al. (1998) — *Biological Sequence Analysis.* The classic on probabilistic models and HMMs.
- Felsenstein (2004) — *Inferring Phylogenies.*
- Burkowski (2008) — *Structural Bioinformatics: An Algorithmic Approach.* Structure, dynamic programming, RNA secondary structure, superposition, machine learning.
- Gagniuc (2021) — *Algorithms in Bioinformatics: Theory and Implementation.* Alignment, motifs, Markov chains.
- Hastie et al. (2009) — *The Elements of Statistical Learning.*

C.2 Keeping up to date

- Proceedings of the most recent **ISMB** and **ECCB** conferences (published as supplements of *Bioinformatics*).
- The annual *Nucleic Acids Research* **Database** and **Web Server** issues.
- Seminars and dissertations at DBB and SciLifeLab.

C.3 Collections of papers

C.3.1 Classical papers in bioinformatics

These papers were read in the PhD course *Classical papers in bioinformatics* (SciLifeLab, Arne Elofsson and Måns Magnusson). Each made a fundamental contribution or changed how the field thinks. For each paper, ask: what was new about it, how did it change the field, and what has happened in the field since?

Protein structure and folding

- Levinthal (1966) — molecular model building by computer; the origin of Levinthal's paradox (Section 9.5).

- Chou and Fasman (1974a); Chou and Fasman (1974b) — the Chou–Fasman secondary-structure propensities and prediction method (Section 9.8).
- Chothia et al. (1977) — how α-helices and β-sheets pack against each other.
- McCammon et al. (1977) — the first molecular dynamics simulation of a protein (Section 9.11).
- Dalal et al. (1997) — "protein alchemy": changing a β-sheet protein into an α-helical one by changing half of its sequence.
- Bradley et al. (2005) — Rosetta reaches high-resolution de novo structure prediction for small proteins (Section 9.8).
- Meng et al. (1992) — automated docking with grid-based energy evaluation (DOCK).

Membrane and signal-peptide prediction

- Heijne (1986) — weight-matrix prediction of signal-peptide cleavage sites.
- Nielsen et al. (1997) — SignalP: neural networks for signal peptides (3).
- Heijne (1992) — hydrophobicity analysis and the positive-inside rule (Section 9.10).
- Krogh et al. (2001) — TMHMM: a hidden Markov model for transmembrane topology (Section 1.6.4).

Phylogeny

- Fitch (1971) — parsimony: minimum change on a given tree (Section 4.3).
- Felsenstein (1981) — maximum-likelihood trees from DNA sequences (Section 4.2).
- Huelsenbeck et al. (2001) — Bayesian inference of phylogeny (Section 4.3).

Sequencing

- Sanger, Nicklen, et al. (1977) — DNA sequencing with chain-terminating inhibitors (Section 5.2).
- Li and Durbin (2009) — BWA: short-read alignment with the Burrows–Wheeler transform (Section 1.4).

D Further explanations

For each main section of the book, open-access reviews, tutorials and textbook chapters that explain the topic in more depth than this guide. They were selected for clarity, not for completeness, and all can be read without a subscription. Each entry gives the title, a DOI or web link, the licence and, where available, the abstract. In the PDF version, one or two particularly useful papers per chapter are reproduced in full at the end of the chapter's section, under their original licences.

D.1 Sequence searching

Further explanations for Chapter 1.

D.1.1 Homology and sequence similarity

- Glover et al. (2019) — A community review of how orthologues and paralogues are defined, inferred and benchmarked. Useful background for the distinction between homology, orthology and paralogy made in this section. *(review; CC BY 4.0)*

  Advances and Applications in the Quest for Orthologs. Glover et al. (2019). *Molecular Biology and Evolution*. doi:10.1093/molbev/msz150

  > Gene families evolve by the processes of speciation (creating orthologs), gene duplication (paralogs), and horizontal gene transfer (xenologs), in addition to sequence divergence and gene loss. Orthologs in particular play an essential role in comparative genomics and phylogenomic analyses. With the continued sequencing of organisms across the tree of life, the data are available to reconstruct the unique evolutionary histories of tens of thousands of gene families. Accurate reconstruction of these histories, however, is a challenging computational problem, and the focus of the Quest for Orthologs Consortium. We review the recent advances and outstanding challenges in this field, as revealed at a symposium and meeting held at the University of Southern California in 2017. Key advances have been made both at the level of orthology algorithm development and with respect to coordination across the community of algorithm developers and orthology end-users. Applications spanned a broad range, including gene function prediction, phylostratigraphy, genome evolution, and phylogenomics. The meetings highlighted the increasing use of meta-analyses integrating results from multiple different algorithms, and discussed ongoing challenges in orthology inference as well as the next steps toward improvement and integration of orthology resources.

D.1.2 Dynamic programming and sequence alignment

- Kellis et al. (2020d) — Lecture-note style chapter (based on MIT course 6.047) that formulates the alignment problem and derives the dynamic-programming solution and the Needleman–Wunsch algorithm step by step, with a short look at multiple alignment. *(book chapter; CC BY-NC-SA 4.0)*

  Sequence Alignment and Dynamic Programming (Chapter 2 of Computational Biology: Genomes, Networks, and Evolution). Kellis et al. (2020). *LibreTexts (MIT open textbook)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/02%3A_Sequence_Alignment_and_Dynamic_Programming

D.1.3 Heuristic search methods

- Kerfeld and Scott (2011) — A short educational article that explains BLAST scores and E-values and how they are used to argue for evolutionary relationships. A gentle introduction to the alignment statistics discussed in this section. *(tutorial; CC BY 4.0)*

  Using BLAST to Teach "E-value-tionary" Concepts. Kerfeld & Scott (2011). *PLoS Biology*. doi:10.1371/journal.pbio.1001014

  > BLAST (Basic Local Alignment Search Tool; [1],[2]) is the key bioinformatic tool for sequence comparison and retrieval from databases. BLAST is often the first step in using sequence-based information to design experiments and contextualize experimental results. Its speed and ease of use help account for this, with experiments requiring simply submitting a sequence of interest (the query sequence) and waiting a few seconds. At the same time, by rendering thinking unnecessary, BLAST is often used suboptimally, with many experienced researchers simply using the default parameters because they do not know how to manipulate them or accepting results with little understanding of their full meaning (or lack thereof). Recognition of the importance of BLAST to modern life sciences has led to its use in many biology courses, even at the high school level, to introduce students to bioinformatics applications in the life sciences. Concepts of molecular evolution (e.g., gene duplication and divergence; orthologs versus paralogs) are quite abstract and are best communicated with examples (Box 1). It is possible to use case studies from the literature, but the abundance of sequence data present in public databases raises the far more attractive possibility of using searches tailored to a particular course, or, better yet, allowing the students to choose their own examples. Box 1. Concepts at a Glance Leads into biological sciences Molecular evolution (e.g., homologs, paralogs, orthologs) Gene alignment Domain structure of proteins Biochemical nature of amino acids: "frequent" and "infrequent" substitutions Conserved versus divergent regions of genes Leads into math Amino acid identity matrices Quantification of sequence identity and similarity Leads into information technology Database queries and their tradeoffs (speed versus completeness) Large dataset management Less obviously, another benefit of teaching students how the BLAST algorithm works is that it provides an opportunity to illustrate how mathematics functions as a language of biology. For example, given that BLAST has been designed to retrieve homologs, there are several steps in the BLAST program that incorporate molecular evolution concepts to maximize the possibility of finding sequences with a shared evolutionary history. More basically, understanding the steps in the calculation of an E-value provides an opportunity to show the relationship between how the algorithm works and fundamental principles of biochemistry and evolution. Here, we provide an approach to teaching the basics of BLAST to students in order to emphasize how the algorithm translates fundamental biological principles into numerical terms culminating in an E value. Acquiring a feel for the algorithm and exploring genomic sequence data has the potential to inform a student's grasp of biomedical, biochemical, and biogeochemical concepts, presenting an excellent opportunity for multidisciplinary integration.

- Kellis et al. (2020b) — Covers global, local and semi-global alignment, linear-time exact string matching and its preprocessing, the BLAST algorithm and the probabilistic foundations of alignment scores. *(book chapter; CC BY-NC-SA 4.0)*

  Rapid Sequence Alignment and Database Search (Chapter 3 of Computational Biology: Genomes, Networks, and Evolution). Kellis et al. (2020). *LibreTexts (MIT open textbook)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/03%3A_Rapid_Sequence_Alignment_and_Database_Search

- Alser et al. (2021) — A broad review of read-mapping algorithms, including hash tables, suffix arrays and the FM-index/Burrows–Wheeler transform. Useful for the exact-matching part of this section. *(review; CC BY 4.0)*

  Technology dictates algorithms: recent developments in read alignment. Alser et al. (2021). *Genome Biology*. doi:10.1186/s13059-021-02443-7

  > Aligning sequencing reads onto a reference is an essential step of the majority of genomic analysis pipelines. Computational algorithms for read alignment have evolved in accordance with technological advances, leading to today's diverse array of alignment methods. We provide a systematic survey of algorithmic foundations and methodologies across 107 alignment methods, for both short and long reads. We provide a rigorous experimental evaluation of 11 read aligners to demonstrate the effect of these

underlying algorithms on speed and efficiency of read alignment. We discuss how general alignment algorithms have been tailored to the specific needs of various domains in biology.

D.1.4 Alignment-free sequence comparison

- Zielezinski et al. (2019) — A community benchmark of many alignment-free methods (k-mer based, MinHash and others) on protein, gene-tree, genome and horizontal-transfer tasks. Shows where these methods work well and where they do not. *(review; CC BY 4.0)*

  Benchmarking of alignment-free sequence comparison methods. Zielezinski et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1755-7

  > Background Alignment-free (AF) sequence comparison is attracting persistent interest driven by data-intensive applications. Hence, many AF procedures have been proposed in recent years, but a lack of a clearly defined benchmarking consensus hampers their performance assessment. Results Here, we present a community resource (http://afproject.org) to establish standards for comparing alignment-free approaches across different areas of sequence-based research. We characterize 74 AF methods available in 24 software tools for five research applications, namely, protein sequence classification, gene tree inference, regulatory element detection, genome-based phylogenetic inference, and reconstruction of species trees under horizontal gene transfer and recombination events. Conclusion The interactive web service allows researchers to explore the performance of alignment-free tools relevant to their data types and analytical goals. It also allows method developers to assess their own algorithms and compare them with current state-of-the-art tools, accelerating the development of new, more accurate AF solutions.

D.1.5 Profiles and multiple sequence alignments

- Notredame (2007) — A concise review of progressive, consistency-based and iterative multiple alignment methods and of how alignment accuracy is benchmarked. It covers the main ideas behind Clustal, T-Coffee, MUSCLE and related tools, though not the most recent methods. *(review; CC BY 4.0)*

  Recent Evolutions of Multiple Sequence Alignment Algorithms. Notredame (2007). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.0030123

  > An ever-increasing number of biological modeling methods depend on the assembly of an accurate multiple sequence alignment (MSA). These include phylogenetic trees, profiles, and structure prediction. Assembling a suitable MSA is not, however, a trivial task, and none of the existing methods have yet managed to deliver biologically perfect MSAs. Many of the algorithms published these last years have been extensively described [1–3], and this review focuses only on the latest developments, including meta-methods and template-based alignment techniques.

- Yoon (2009) — A tutorial-style review of HMMs, the Viterbi, forward and backward algorithms, parameter estimation and profile HMMs for sequence families. Fits the HMM and profile-HMM parts of this section. *(tutorial; CC BY (Europe PMC))*

  Hidden Markov Models and their Applications in Biological Sequence Analysis. Yoon (2009). *Current Genomics*. doi:10.2174/138920209789177575

  > Hidden Markov models (HMMs) have been extensively used in biological sequence analysis. In this paper, we give a tutorial review of HMMs and their applications in a variety of problems in molecular biology. We especially focus on three types of HMMs: the profile-HMMs, pair-HMMs, and context-sensitive HMMs. We show how these HMMs can be used to solve various sequence analysis problems, such as pairwise and multiple sequence alignments, gene annotation, classification, similarity search, and many others.

- Kellis et al. (2020a) — Introduces Markov chains and HMMs with a casino example and describes the main HMM algorithms, close to the worked example used in this section. *(book chapter; CC BY-NC-SA 4.0)*

  Hidden Markov Models I (Chapter 7 of Computational Biology: Genomes, Networks, and Evolution). Kellis et al. (2020). *LibreTexts (MIT open textbook)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/07%3A_Hidden_Markov_Models_I

D.1.6 RNA homology search

- Kellis et al. (2020c) — Explains RNA secondary structure, dynamic-programming approaches to the folding problem, the evolution of RNA structure and a probabilistic approach to folding. *(book chapter; CC BY-NC-SA 4.0)*

  RNA Folding (Chapter 10 of Computational Biology: Genomes, Networks, and Evolution). Kellis et al. (2020). *LibreTexts (MIT open textbook)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/10%3A_RNA_Folding

- Zhao et al. (2021) — Reviews thermodynamic, comparative and machine-learning approaches to RNA secondary structure prediction. Background for the structure-prediction part of this section rather than for homology search itself. *(review; CC BY 4.0)*

  Review of machine learning methods for RNA secondary structure prediction. Zhao et al. (2021). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1009291

  > Secondary structure plays an important role in determining the function of noncoding RNAs. Hence, identifying RNA secondary structures is of great value to research. Computational prediction is a mainstream approach for predicting RNA secondary structure. Unfortunately, even though new methods have been proposed over the past 40 years, the performance of computational prediction methods has stagnated in the last decade. Recently, with the increasing availability of RNA structure data, new methods based on machine learning (ML) technologies, especially deep learning, have alleviated the issue. In this review, we provide a comprehensive overview of RNA secondary structure prediction methods based on ML technologies and a tabularized summary of the most important methods in this field. The current pending challenges in the field of RNA secondary structure prediction and future trends are also discussed.

D.1.7 Databases and domains

- Das et al. (2015) — A short review, based on CATH, of how domain superfamilies diversify in structure and function. Useful for the structural classification of domains. *(review; CC BY 4.0)*

  Diversity in protein domain superfamilies. Das et al. (2015). *Current Opinion in Genetics & Development*. doi:10.1016/j.gde.2015.09.005

  > Whilst ~93% of domain superfamilies appear to be relatively structurally and functionally conserved based on the available data from the CATH-Gene3D domain classification resource, the remainder are much more diverse. In this review, we consider how domains in some of the most ubiquitous and promiscuous superfamilies have evolved, in particular the plasticity in their functional sites and surfaces which expands the repertoire of molecules they interact with and actions performed on them. To what extent can we identify a core function for these superfamilies which would allow us to develop a 'domain grammar of function' whereby a protein's biological role can be proposed from its constituent domains? Clearly the first step is to understand the extent to which these components vary and how changes in their molecular make-up modifies function.

- Forslund et al. (2019) — Explains what domains and domain architectures are, how they are annotated with resources such as Pfam and InterPro, and how architectures evolve by fusion, fission and shuffling. *(book chapter; CC BY 4.0)*

  Evolution of Protein Domain Architectures. Forslund et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_15

  > This chapter reviews current research on how protein domain architectures evolve. We begin by summarizing work on the phylogenetic distribution of proteins, as this will directly impact which domain architectures can be formed in different species. Studies relating domain family size to occurrence have shown that they generally follow power law distributions, both within genomes and larger evolutionary groups. These findings were subsequently extended to multi-domain architectures. Genome evolution models that have been suggested to explain the shape of these distributions are reviewed, as well as evidence for selective pressure to expand certain domain families more than others. Each domain has an intrinsic combinatorial propensity, and the effects of this have been studied using measures of domain versatility or promiscuity. Next, we study the principles of protein domain architecture evolution and how these have been inferred from distributions of extant domain arrangements. Following this, we review inferences of ancestral domain architecture and the conclusions concerning domain architecture evolution mechanisms that can be drawn from these. Finally, we examine whether all known cases of a given domain architecture can be assumed to have a single common origin (monophyly) or have evolved convergently (polyphyly). We end by a discussion of some available tools for computational analysis or exploitation of protein domain architectures and their evolution.

- Wang et al. (2021) — An overview of sequence- and structure-based methods and databases for identifying protein domains. Link only because of the ND licence. *(review; CC BY-NC-ND 4.0 (link only))*

  Protein domain identification methods and online resources. Wang et al. (2021). *Computational and Structural Biotechnology Journal*. doi:10.1016/j.csbj.2021.01.041

  > Protein domains are the basic units of proteins that can fold, function, and evolve independently. Knowledge of protein domains is critical for protein classification, understanding their biological functions, annotating their evolutionary mechanisms and protein design. Thus, over the past two decades, a number of protein domain identification approaches have been developed, and a variety of protein domain databases have also been constructed. This review divides protein domain prediction methods into two categories, namely sequence-based and structure-based. These methods are introduced in detail, and their advantages and limitations are compared. Furthermore, this review also provides a comprehensive overview of popular online protein domain sequence and structure databases. Finally, we discuss potential improvements of these prediction methods.

D.1.8 Protein language models

- Chandra et al. (2023) — Explains the transformer architecture and its use in protein language models for predicting protein properties. A readable introduction for students without a machine-learning background. *(review; CC BY 4.0)*

  Transformer-based deep learning for predicting protein properties in the life sciences. Chandra et al. (2023). *eLife*. doi:10.7554/elife.82819

  > Recent developments in deep learning, coupled with an increasing number of sequenced proteins, have led to a breakthrough in life science applications, in particular in protein property prediction. There is hope that deep learning can close the gap between the number of sequenced proteins and proteins with known properties based on lab experiments. Language models from the field of natural language processing have gained popularity for protein property predictions and have led to a new computational revolution in biology, where old prediction results are being improved regularly. Such models can learn useful multipurpose representations of proteins from large open repositories of protein sequences and can be used, for instance, to predict protein properties. The field of natural language processing is growing quickly because of developments in a class of models based on a particular model-the Transformer model. We review recent developments and the use of large-scale Transformer models in applications

> for predicting protein characteristics and how such models can be used to predict, for example, post-translational modifications. We review shortcomings of other deep learning models and explain how the Transformer models have quickly proven to be a very promising way to unravel information hidden in the sequences of amino acids.

- Valentini et al. (2023) — Surveys protein language models, their training and their uses in structure and function prediction and protein design. Gives broader context for this section. *(review; CC BY 4.0)*

  The promises of large language models for protein design and modeling. Valentini et al. (2023). *Frontiers in Bioinformatics*. doi:10.3389/fbinf.2023.1304099

  > The recent breakthroughs of Large Language Models (LLMs) in the context of natural language processing have opened the way to significant advances in protein research. Indeed, the relationships between human natural language and the "language of proteins" invite the application and adaptation of LLMs to protein modelling and design. Considering the impressive results of GPT-4 and other recently developed LLMs in processing, generating and translating human languages, we anticipate analogous results with the language of proteins. Indeed, protein language models have been already trained to accurately predict protein properties, generate novel functionally characterized proteins, achieving state-of-the-art results. In this paper we discuss the promises and the open challenges raised by this novel and exciting research area, and we propose our perspective on how LLMs will affect protein modeling and design.

- Ofer et al. (2021) — Explains the analogy between natural language processing and protein sequences, from word embeddings to transformers. Link only because of the ND licence. *(review; CC BY-NC-ND 4.0 (link only))*

  The language of proteins: NLP, machine learning & protein sequences. Ofer et al. (2021). *Computational and Structural Biotechnology Journal*. doi:10.1016/j.csbj.2021.03.022

  > Natural language processing (NLP) is a field of computer science concerned with automated text and language analysis. In recent years, following a series of breakthroughs in deep and machine learning, NLP methods have shown overwhelming progress. Here, we review the success, promise and pitfalls of applying NLP algorithms to the study of proteins. Proteins, which can be represented as strings of amino-acid letters, are a natural fit to many NLP methods. We explore the conceptual similarities and differences between proteins and language, and review a range of protein-related tasks amenable to machine learning. We present methods for encoding the information of proteins as text and analyzing it with NLP methods, reviewing classic concepts such as bag-of-words, k-mers/n-grams and text search, as well as modern techniques such as word embedding, contextualized embedding, deep learning and neural language models. In particular, we focus on recent innovations such as masked language modeling, self-supervised learning and attention-based models. Finally, we discuss trends and challenges in the intersection of NLP and protein research.

D.1.9 Benchmarking search methods

- Weber et al. (2019) — General guidelines for designing fair benchmarks: choosing methods and reference datasets, defining the ground truth and choosing evaluation metrics. These apply directly to the benchmarking of homology search methods. *(review; CC BY 4.0)*

  Essential guidelines for computational method benchmarking. Weber et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1738-8

  > In computational biology and other sciences, researchers are frequently faced with a choice between several computational methods for performing data analyses. Benchmarking studies aim to rigorously compare the performance of different methods using well-characterized benchmark datasets, to determine the strengths of each method or to provide recommendations regarding suitable choices of methods for an analysis. However, benchmarking studies must be carefully designed and implemented to provide accurate, unbiased, and informative results. Here, we summarize key practical guidelines and recommendations for performing high-quality benchmarking analyses, based on our experiences in computational biology.

D.1.10 Whole-genome alignment and conservation

- Armstrong et al. (2019) — Explains pairwise and multiple whole-genome alignment, the Cactus approach, and how alignments are used to annotate genes and conserved elements across species. *(review; licence not stated)*

  Whole-Genome Alignment and Comparative Annotation. Armstrong et al. (2019). *Annual Review of Animal Biosciences*. doi:10.1146/annurev-animal-020518-115005

  > Rapidly improving sequencing technology coupled with computational developments in sequence assembly are making reference-quality genome assembly economical. Hundreds of vertebrate genome assemblies are now publicly available, and projects are being proposed to sequence thousands of additional species in the next few years. Such dense sampling of the tree of life should give an unprecedented new understanding of evolution and allow a detailed determination of the events that led to the wealth of biodiversity around us. To gain this knowledge, these new genomes must be compared through genome alignment (at the sequence level) and comparative annotation (at the gene level). However, different alignment and annotation methods have different characteristics; before starting a comparative genomics analysis, it is important to understand the nature of, and biases and limitations inherent in, the chosen methods. This review is intended to act as a technical but high-level overview of the field that should provide this understanding. We briefly survey the state of the genome alignment and comparative annotation fields and potential future directions for these fields in a new, large-scale era of comparative genomics.

The following open-access paper(s) are reproduced under their original licences.

OPEN ACCESS Freely available online

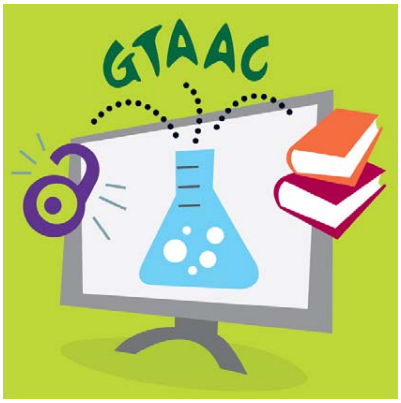
PLOS BIOLOGY

Education

Using BLAST to Teach "E-value-tionary" Concepts

Cheryl A. Kerfeld[1,2]*, Kathleen M. Scott[3]

1 Joint Genome Institute, Walnut Creek, California, United States of America, **2** Department of Plant and Microbial Biology, University of California, Berkeley, California, United States of America, **3** Integrative Biology Department, University of South Florida, Tampa, Florida, United States of America



Introduction

BLAST (Basic Local Alignment Search Tool; [1,2]) is the key bioinformatic tool for sequence comparison and retrieval from databases. BLAST is often the first step in using sequence-based information to design experiments and contextualize experimental results. Its speed and ease of use help account for this, with experiments requiring simply submitting a sequence of interest (the query sequence) and waiting a few seconds. At the same time, by rendering thinking unnecessary, BLAST is often used suboptimally, with many experienced researchers simply using the default parameters because they do not know how to manipulate them or accepting results with little understanding of their full meaning (or lack thereof).

Recognition of the importance of BLAST to modern life sciences has led to its use in many biology courses, even at the high school level, to introduce students to bioinformatics applications in the life sciences. Concepts of molecular evolution (e.g., gene duplication and divergence; orthologs versus paralogs) are quite abstract and are best communicated with examples (Box 1). It is possible to use case studies from the literature, but the abundance of sequence data present in public databases raises the far more attractive possibility of using searches tailored to a particular course, or, better yet, allowing the students to choose their own examples.

Less obviously, another benefit of teaching students how the BLAST algorithm works is that it provides an opportunity to illustrate how mathematics functions as a language of biology. For example, given that BLAST has been designed to retrieve homologs, there are several steps in the BLAST program that incorporate molecular evolution concepts to maximize the possibility of finding sequences with a shared evolutionary history. More basically, understanding the steps in the calculation of an E-value provides an opportunity to show the relationship between how the algorithm works and fundamental principles of biochemistry and evolution. Here, we provide an approach to teaching the basics of BLAST to students in order to emphasize how the algorithm translates fundamental biological principles into numerical terms culminating in an E value. Acquiring a feel for the algorithm and exploring genomic sequence data has the potential to inform a student's grasp of biomedical, biochemical, and biogeochemical concepts, presenting an excellent opportunity for multidisciplinary integration.

Explaining the Relationship between Aligning Sequences and Evolutionary Biology

To begin a BLAST search of a database, the user provides a query sequence, which is a nucleotide (nt) or amino acid (aa) sequence for which they are interested in finding homologs. The BLAST algorithm begins by fragmenting the sequence into "words" (e.g., from the National Center for Biotechnology Information (NCBI) interface, 16–256 nt, or 2–3 aa), and, from each word, creating a set of acceptable "synonyms" that represent possible changes in sequence due to mutation (Figure 1). Words and their synonyms are scored with respect to how well they match the query sequence, based on substitution matrices (see below) from curated alignments of gene/protein families (e.g., BLOcks of Amino Acid SUbstitution Matrix [BLOSUM]; [3,4]). The words that match sufficiently well to have a score above a set threshold value are carried forward to compare to all the sequences in the database being searched for homologs. All sequences in the database are then scanned for the presence of these words; sequences carrying two matches within a preset distance from each other (which suggests a conserved region shared by both query and subject sequences) are set aside until the entire database has been scanned. This "short list" of subject sequences is then carried

The Education Series features noteworthy, innovative open-education programs to enhance understanding of biology.

Citation: Kerfeld CA, Scott KM (2011) Using BLAST to Teach "E-value-tionary" Concepts. PLoS Biol 9(2): e1001014. doi:10.1371/journal.pbio.1001014

Series Editor: Cheryl A. Kerfeld, University of California Berkeley/JGI, United States of America

Published February 1, 2011

Copyright: © 2011 Kerfeld, Scott. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Curriculum development activities of CAK and KMS are supported in part by a USDA Higher Education Challenge Grant, MOE-2008-02036, as well as NSF MCB-0643713 to KMS. The work of CAK at the JGI is under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: aa, amino acid; BLAST, Basic Local Alignment Search Tool; NCBI, National Center for Biotechnology Information; nt, nucleotide.

* E-mail: CKerfeld@lbl.gov

Box 1. Concepts at a Glance

Leads into biological sciences

- Molecular evolution (e.g., homologs, paralogs, orthologs)
- Gene alignment
- Domain structure of proteins
- Biochemical nature of amino acids: "frequent" and "infrequent" substitutions
- Conserved versus divergent regions of genes

Leads into math

- Amino acid identity matrices
- Quantification of sequence identity and similarity

Leads into information technology

- Database queries and their tradeoffs (speed versus completeness)
- Large dataset management

1. Generate words from sequence above threshold (e.g. T=11)

Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS
VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

SWVSQASFTPPGIM → SWV WVS VSQ SQA QAS ASF SFT ...

Selection of words scoring above threshold (for word SWV):

Substitution Matrix*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G		6	-4	-2	-3	0	-2	-2	-3
I			4	-3	0	-2	-1	-3	3
K				5	-3	0	-1	-3	-2
F					6	-2	-2	1	-1
S						4	1	-3	-2
T							5	-2	0
W								11	-3
V									4

*A portion of the BLOSUM 62 matrix

SWV (4+11+4 = 19)
SWI (4+11+3 = 18)
TWV (1+11+4 = 16)
GWV (0+11+4 = 15)
KWV (0+11+4 = 15)
SWS (4+11-2 = 13)

Synonyms above threshold 11... (others not shown)

SFV (4+1+4 = 9)
SRV (4-3+4 = 5)

Synonyms below threshold 11... (others not shown)

2. Search the database for words matching those generated

3. Extend matching hits in both directions

```
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNK
    .| ||||||||| |||: :|:   |   | ||
..TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD
```

Word match from Step 1 — Extension until score drops

4. Generate alignment and calculate statistics

```
>ref|YP_002482587.1| flavin reductase domain protein FMN-binding [Cyanothece sp.
PCC 7425]
 gb|ACL44226.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 Score =  176 bits (446),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query  1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +    H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                 +L SWVSQASF PPG+ +A+  E  A GL     AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS    GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

Figure 1. Steps in the BLAST algorithm.
doi:10.1371/journal.pbio.1001014.g001

forward by extending the alignment outward from the words to determine whether the match between the query and subject sequences extends beyond the local match between the subject sequence and the word. Initial "rough" alignments are extended without gaps to verify that the sequences match beyond the word hits. If the threshold score for the "ungapped" alignment is high enough that it suggests that the two sequences are indeed homologs, a second alignment is undertaken in which gaps (see below) are allowed to optimize the alignment. The sequences retrieved after these steps are referred to as the "subject" sequences.

The extent of the sequence similarity between the subject and query sequences is reported as a raw score, S. A pairwise alignment between a subject sequence and query sequence in which the two have a high degree of sequence similarity will have a high S value. S is calculated from the alignment of the two sequences by adding the scores from each pair ij of residues in the alignment:

$$S = \left(\sum M_{ij}\right) - cO - dG$$

in which M is the score from a similarity matrix (e.g., BLOSUM62; more on this below) for a particular pair of amino acids, c is the number of gaps, O is the penalty for the existence of a gap, d is the total length of the gaps, and G is the per-residue penalty for extending the gap.

Gaps are spaces introduced into either the query or subject sequence to optimize alignments. In terms of the underlying biology of the two aligned sequences, gaps correspond to regions in which there is an insertion or a deletion in one of the amino acid sequences relative to the other. In a protein BLAST, the default setting for O is 11, while for G it is 1. It is useful to ask students to consider why the gap existence penalty is larger than the gap extension penalty. The answer is that it is assumed that a gap corresponds to either an insertion or deletion event, regardless of the size of the gap. Hence, a larger penalty is imposed for the existence of the gap whereas the length of that insertion or deletion is relatively less significant. Furthermore, these insertion and deletion events are likely rare and so more heavily penalized. If they were not, and the gap opening penalty correspondingly lowered, numerous gaps would be included in alignments; taken to the extreme, this would allow any two sequences to be aligned.

Alignments are extended position by position with concomitant scoring until

the match falls below a threshold score, at which point it is terminated. It is important to note that alignment between the query and subject sequences does not have to cover the full length of the two sequences. Therefore, retrieved subject sequences commonly align with only a portion of the query sequence—it is this "local" rather than global quality that is more than nominally BLAST's strength. Underscoring this distinction for students provides another opportunity to discuss protein evolution. In considering their results, they may need to evaluate if they have found homology in one domain of a multidomain protein. BLAST provides a list of potential homologs to the query sequence that is based on regions of sequence similarity. Matches that occur over limited regions of these sequences provide an illustration of both the modular nature of protein structure (domains), as well as the modular nature of protein evolution (new domain combinations). Protein domains with a common ancestry are shared among many protein families whose members have divergent functions. For example, flavin reductase domains occur in the many enzymes that transfer electrons from a reduced compound (e.g., NADH) to FMN or FAD. If two amino acid sequences have a flavin reductase domain but do not share other domains with common ancestry, the two sequences will align with a high score at the flavin reductase domain, but will not align meaningfully or significantly elsewhere (Figure 1).

Sometimes it is helpful to "mask" parts of the query sequence to prevent them from being aligned with subject sequences. Masking is helpful when the query sequence has "low-complexity" regions, such as stretches of small hydrophobic amino acids that are commonly present in transmembrane helices of integral membrane proteins. Because these features arose from convergent evolution, and their inclusion in BLAST searches could result in spurious hits, it is best to set the BLAST search parameters to eliminate these sorts of regions from word generation, as well as alignment scoring.

Why Aren't Raw Scores Enough? The Calculation of the E-Value

Because one has the option of using different parameters (e.g., matrices) in different BLAST experiments, it is ideal to report results in such a way as to be able to compare alignments made with different scoring matrices or gap penalties. To do this, S' values (bit scores) are calculated:

$$S' = (\lambda S - \ln K)/\ln 2$$

in which λ and K reflect the matrices and penalties used. If analyses were to stop here, one would have a list of sequences sorted by bit scores that would reflect the degree of similarity to the query sequence. The question of whether the sequences were homologs to the query sequence would not yet be directly addressed: how high does a bit score have to be to suggest shared ancestry? Larger databases are more likely to include sequences with matches to the query that are due to chance, not homology.

To address this issue, E-values are calculated from bit scores as

$$E = (n \times m)/\left(2^{S'}\right)$$

in which n is the total number of residues (amino or nucleic acids) in the database, and m is the length of the query sequence. E-values are the number of subject sequences that can be expected to be retrieved from the database that have a bit score equal to or greater than the one calculated from the alignment of the query and subject sequence, based on chance alone. E-values for subject sequences that are very similar to the query sequence will be quite small, and are widely used as a means to assess the confidence with which one should claim the subject sequence(s) and the query sequence as homologs. Collectively, the parameters for a BLAST search are given at the bottom of the output; going through these with students provides another opportunity to reinforce these concepts and explain why an E-value for BLAST searches using the same protein sequence will change over time.

Substitution Matrices and Protein Biochemistry

Substitution matrices are critical at two steps in the BLAST process: 1) in the evaluation of "words" with which to tag subject sequences for further scrutiny, and 2) in the extension and evaluation of alignments between the query sequence and a subject sequence. These matrices are used to calculate scores for alignments, which are used to gauge the strength of the match between two sequences. Scores are calculated by comparing residues at each position in the alignment. This series of scores is summed over the length of the alignment and used to calculate the raw score.

It is intuitive that having an exact match in residues at many positions in an alignment should result in a high alignment score; it is less obvious how these matches should be weighted among different residues (is a leucine–leucine match as informative as a tryptophan–tryptophan match?), as well as how severely residue substitutions should be penalized (should a leucine–isoleucine sub-

Box 2. Teaching Tools

- Powerpoint slides for teaching BLAST (Supporting Information S1)

Online resources

- National Center for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/guide/
  - BLAST interface to query vast protein and nucleotide sequence databases, and protein structure databases
  - Links to PubMed
  - Tutorials for resources
- Bioinformatics tutorials at the European Bioinformatics Institute (EBI): http://www.ebi.ac.uk/2can/tutorials/index.html
  - BLAST background http://www.ebi.ac.uk/2can/tutorials/protein/blast.html
- Integrated Microbial Genomes (IMG): http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
  - BLAST interface to query sequenced microbial genomes
  - Simple tools for genome: genome comparison
- PFAM: http://pfam.sanger.ac.uk/
  - BLAST interface to query curated protein families
  - Comprehensive background information and literature about protein families

Box 3. Evaluation Tools

Questions to ask before the lesson

- What is a homolog?
- How can you tell if two genes are likely to share a common ancestor?
- How can you tell if two proteins share a common domain?
- Is the degree of harm of a mutation in a gene sensitive to its position in a gene?

Questions to ask after the lesson

- Can you ever prove that two genes share a common ancestor? Why or why not?
- How could you build a strong argument that two genes have the same or similar function in two different organisms?
- Can you, in your own words, describe the process of a BLAST search?
- Which steps in a BLAST search help to make it happen so quickly?

stitution be penalized to the same degree as a leucine–tryptophan substitution?).

Substitution matrices were created to resolve these "gray areas"; they consist of factors to use to weigh residue matches and substitutions. BLOSUM matrices are commonly used to compare amino acid sequences and were constructed from curated alignments of amino acid sequences [4]. The alignments were scanned for positions at which sequences diverged, and the frequencies of substitutions for each amino acid were noted; some amino acid substitutions were frequently observed (e.g., due to chemical similarity between residues) and therefore should be penalized less stringently than those rarely observed. Amino acid "commonness" was also considered; exact matches between more pedestrian amino acids (e.g., leucine) are not scored as highly as more "exotic" ones (e.g., tryptophan).

Different levels of overall sequence divergence also affect amino acid substitution frequencies (e.g., fewer substitutions are expected for less divergent sequences; therefore, the substitutions that do occur should be penalized more heavily). To account for this effect, BLOSUM matrices were constructed from alignments with different degrees of divergence. BLOSUM62 was created from an alignment of sequences with at least 62% sequence identity, while BLOSUM80 used an alignment with 80% [4].

Examining and comparing substitution matrices with students provides an opportunity to consider not only the consequences of residue substitution but also the reason(s) for the different substitution penalties, allowing them to apply what they know about amino acid chemistry. When comparing two BLOSUM matrices, students will need to consider the types of amino acid replacements expected to accumulate between recently versus deeply diverging sequences.

Meaning and Results

Students (and researchers as well) tend to draw an arbitrary line below which they consider E-values to provide convincing evidence that two sequences are homologs (e.g., $E<0.00001$). It is informative to scrutinize this assumption, and ask the students to consider whether and when more stringent E-values might be appropriate (e.g., to assist in sorting paralogs from orthologs), or when larger E-values do not provide definitive evidence of evolutionary independence (as is the case when two sequences share an ancient ancestor).

It is also informative for the students to discuss what their alignments "mean", and whether the pairwise alignments between their query sequence and the subject sequences "prove" whether the sequences are homologs. Indeed, it can catalyze a larger discussion of whether it is possible to "prove" that two sequences are homologs, and what other approaches (e.g., protein structure, gene context) might be used to strengthen or refute such an assertion.

A meaningful alignment will facilitate the comparison of two sequences with a shared evolutionary history by maximizing the juxtaposition of similar and identical residues. Sequences with a recent shared ancestry will have a high degree of similarity; their alignments will have many identical residues, few substitutions and gaps, and tiny E-values. Conversely, sequences with an ancient common ancestor will be deeply divergent, with few shared sequence identities, many gaps, and larger E-values. Furthermore, an alignment of two sequences can clarify which portions are conserved (e.g., active sites), and which are divergent, which helps cultivate students' understanding of protein structure and function.

In summary, deconstructing the BLAST algorithm and manipulating parameters systematically and evaluating the results with students helps them understand not only what the scores mean but also how to manipulate parameters to optimize their searches. There is a wealth of additional resources available (Box 2; Supporting Information S1 and Text S1). Finally, explicating the algorithm in this way allows students to explore research databases thoughtfully and illustrates the critical connection between mathematics and science, showing how numbers can be used to quantify biological relationships from the level of gene to organism (Box 3). As bioinformatics and databases will increasingly become part of the undergraduate life sciences curriculum, new opportunities are emerging to teach biological principles through their bioinformatic articulations.

Supporting Information

Supporting Information S1 Powerpoint file--Using BLAST To Teach "E-value-tionary" Concepts.
(781 KB PPT)

Text S1 Acquiring a feel for the algorithm: manipulating the parameters of BLAST and other experiments.
(16.8 KB DOCX)

Acknowledgments

The authors gratefully acknowledge helpful conversations with Brad Goodner, Sam Donovan, and Shari Freyermuth. We also thank our students, past and present, who have inspired the writing of this manuscript. We thank Seth Axen for assistance in preparation of the figures for the manuscript.

References

1. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
3. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nature Biotechnol 22: 1035–1036.
4. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89: 10915–10919.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

Recent Evolutions of Multiple Sequence Alignment Algorithms

Cédric Notredame

An ever-increasing number of biological modeling methods depend on the assembly of an accurate multiple sequence alignment (MSA). These include phylogenetic trees, profiles, and structure prediction. Assembling a suitable MSA is not, however, a trivial task, and none of the existing methods have yet managed to deliver biologically perfect MSAs. Many of the algorithms published these last years have been extensively described [1–3], and this review focuses only on the latest developments, including meta-methods and template-based alignment techniques.

The purpose of an MSA algorithm is to assemble alignments reflecting the biological relationship between several sequences. Computing exact MSAs is computationally almost impossible, and in practice approximate algorithms (heuristics) are used to align sequences, by maximizing their similarity. The biological relevance of these MSAs is usually assessed by systematic comparison with established collections of structure-based MSAs ("gold standards"; for review see [4]). Since only a few sequences have known structures, the accuracy measured on the references is merely an estimation of how well a package may fare on standard datasets. Gold standards have had a considerable effect on the evolution of MSA algorithms, refocusing the entire methodological development toward the production of structurally correct alignments. Their use has also coincided with a notable algorithmic harmonization, most MSA packages being now based on the "progressive algorithm" [5]. This greedy heuristic assembly algorithm involves estimating a guide tree (rooted binary tree) from unaligned sequences, and then incorporating the sequences into the MSA with a pairwise alignment algorithm while following the tree topology. The progressive algorithm is often embedded in an iterative loop where the guide tree and the MSA are reestimated until convergence. Most MSA packages reviewed here [6–18] follow this canvas, albeit more or less extensively adapted for improved performances [1–3].

The scoring schemes used by the pairwise alignment algorithm are arguably the most influential component of the progressive algorithm. They can be divided in two categories: matrix- and consistency-based. Matrix-based algorithms such as ClustalW [14], MUSCLE [6], and Kalign [19] use a substitution matrix to assess the cost of matching two symbols or two profiled columns. Although profile statistics can be more or less sophisticated, the score for matching two positions depends only on the considered columns or their immediate surroundings. By contrast, consistency-based schemes incorporate a larger share of information into the evaluation. This result is achieved by using a recipe initially developed for T-Coffee [10] and inspired by Dialign overlapping weights [20]. Its principle is to compile a collection of pairwise global and local alignments (primary library) and to use this collection as a position-specific substitution matrix during a regular progressive alignment. The aim is to deliver a final MSA as consistent as possible with the alignments contained in the library. Many recent packages have built upon this initial framework. For instance, PCMA [15] decreases T-Coffee computational requirements by prealigning closely related sequences. ProbCons [7] uses Bayesian consistency and fills the primary library using the posterior decoding of a pair hidden Markov model. The substitution costs are estimated from this library using Bayesian statistics. MUMMALS [17] combines the ProbCons scoring scheme with the PCMA strategy, while including secondary structure predictions in its pair hidden Markov model. The most accurate flavors of MAFFT [8] (i.e., the GNS and LNS modes) use a T-Coffee–like evaluation. A majority of studies indicate that consistency-based methods are more accurate than their matrix-based counterparts [4], although they typically require an amount of CPU time N times higher than simpler methods (N being the number of sequences). Most of these methods are available online, either as downloadable packages or as online Web services (Table 1).

The wealth of available methods and their increasingly similar accuracies makes it harder than ever to objectively choose one over the others. Consensus methods such as M-Coffee [12] provide an interesting framework to address this problem. M-Coffee is a consensus meta-method based on T-Coffee. Given a sequence dataset, it fills up the library by using various MSA methods to compute alternative alignments. T-Coffee then uses this library to compute a final MSA consistent with the original alignments. When combining eight of the most accurate and distinct MSA packages, M-Coffee produces 67% of the time a better MSA than ProbCons, the best individual method [12]. Aside from its ease of extension, M-Coffee's main advantage is its ability to estimate the local consistency between the final alignment and the combined MSAs (CORE index [21]; Figure 1). This useful index has been shown to be well-correlated with the MSAs' structural correctness [21,22]. M-Coffee is not, however, the ultimate answer to the MSA problem, and its limited performances on remote homologs suggest that

Editor: Fran Lewitter, Whitehead Institute, United States of America

Citation: Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3(8): e123. doi:10.1371/journal.pcbi.0030123

Copyright: © 2007 Cédric Notredame. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: MSA, multiple sequence alignment

Cédric Notredame is with Information Génomique et Structurale, CNRS UPR2589, Institute for Structural Biology and Microbiology, Parc Scientifique de Luminy, Marseille, France. E-mail: cedric.notredame@europe.com

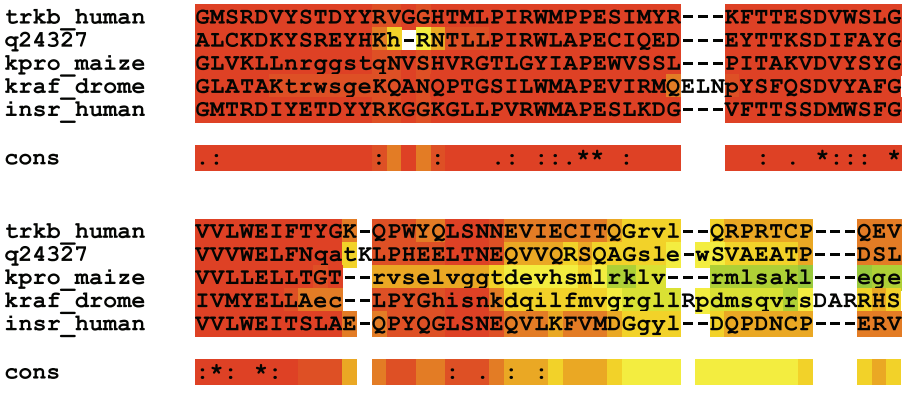
Table 1. Summary of the Methods Described in the Review

Method	Score	Templates	Validation Values		Server
			PreFab	HOMSTRAD	
ClustalW [14]	Matrix	—	61.80 [12]	—	http://www.ebi.ac.uk/clustalw/
Kalign	Matrix	—	63.00 [18]	—	http://msa.cgb.ki.se/
MUSCLE [6]	Matrix	—	68.00 [16]	45.0 [9]	http://www.drive5.com/muscle/
T-Coffee [10]	Consistency	—	69.97 [12]	44.0 [9]	http://www.tcoffee.org/
ProbCons [7]	Consistency	—	70.54 [12]	—	http://probcons.stanford.edu/
MAFFT [8]	Consistency	—	72.20 [12]	—	http://align.genome.jp/mafft/
M-Coffee [12]	Consistency	—	72.91 [12]	—	http://www.tcoffee.org/
MUMMALS [16]	Consistency	—	73.10 [16]	—	http://prodata.swmed.edu/mummals/
DbClustal [24]	Profiles	—	—	—	http://bips.u-strasbg.fr/PipeAlign/
PRALINE [9]	Matrix	Profiles	—	50.2 [9]	http://zeus.cs.vu.nl/programs/pralinewww/
PROMALS [16]	Consistency	Profiles	79.00 [16]	—	http://prodata.swmed.edu/promals/
SPEM [28]	Matrix	Profiles	77.00 [28]	—	http://sparks.informatics.iupui.edu/Softwares-Services_files/spem.htm
Expresso [13]	Consistency	Structures	—	71.9 [11][a]	http://www.tcoffee.org/
T-Lara [29]	Consistency	Structures	—	—	https://www.mi.fu-berlin.de/w/LiSA/

Validation values were compiled from several sources, and selected for comparability. PreFab validations were made using PreFab version 3. HOMSTRAD validations were made on datasets having less than 30% identity. The source of each value is indicated by the accompanying reference citation.
[a]The Expresso value comes from a slightly more demanding subset of HOMSTRAD (HOM39) made of sequences less than 25% identical.
doi:10.1371/journal.pcbi.0030123.t001

further improvement using only sequence information remains an elusive goal. Progress is nonetheless needed, and, at this point, the most promising approach is probably to incorporate within the datasets any information likely to improve the alignments, such as structural and homology data. Template-based alignment methods [13] follow this approach.

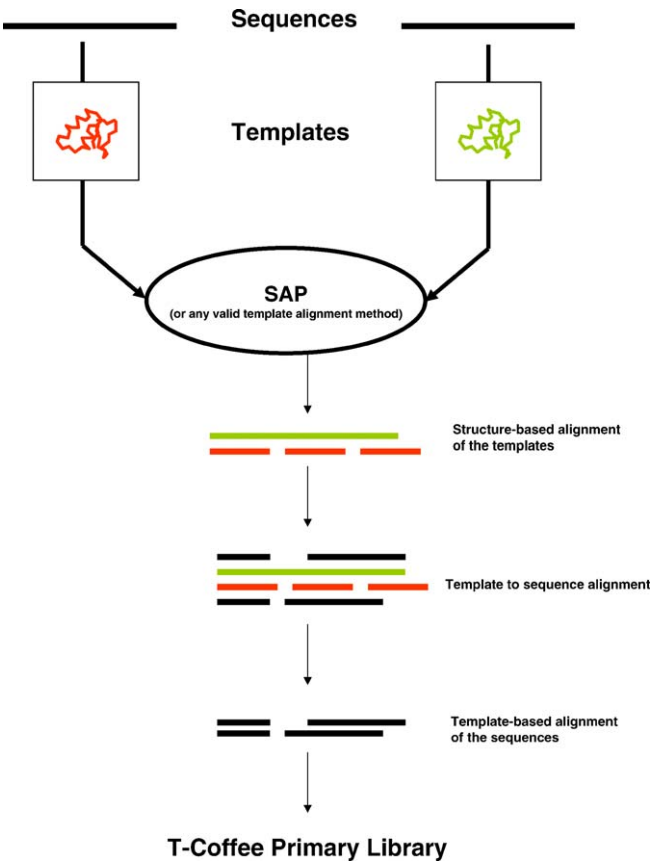
Structural extension was initially described by Taylor [23]. The principle is fairly straightforward (Figure 2) and involves identifying with BLAST a structural template in the Protein Data Bank for each sequence, aligning the templates using a structure superposition method, and mapping the original sequences onto their template's alignment. The resulting sequence alignments are compiled in the primary library and used by a consistency-based method to compute the final MSA. Homology extension was originally introduced in the DbClustal package [24] and works along the same lines, using a profile rather than a structure. PSI-BLAST is used to build a profile for each sequence, and these profiles are used as templates to generate better sequence alignments, thanks to the evolutionary information they contain. The only difference between homology and structure extension is the templates' nature and the associated alignment method. This generic approach can easily be extended to any kind of template. For instance, Expresso [13] uses SAP [25,26] and FUGUE [27] to align structural templates identified by a BLAST against the Protein Data Bank. PROMALS [17], PRALINE [9], and SPEM [28] make a profile–profile alignment with PSI-BLAST profiles used as templates. In PRALINE and PROMALS, the profile can be complemented with a secondary structure prediction in an attempt to improve the alignment accuracy. PROMALS uses ProbCons Bayesian consistency to fill its library with the posterior decoding of a pair hidden Markov model. T-Lara [29] uses



doi:10.1371/journal.pcbi.0030123.g001

Figure 1. Typical Output of M-Coffee

This output was obtained on the kinase1_ref5 BaliBase dataset, by combining MUSCLE, MAFFT, POA, Dialign-T, T-Coffee, ClustalW, PCMA, and ProbCons with M-Coffee. Correctly aligned residues (as judged from the reference) are uppercase; noncorrect ones are lowercase. The color of each residue indicates the agreement of the individual MSAs with respect to the alignment of that specific residue. Red indicates residues aligned in a similar fashion among all the individual MSAs; blue indicates very low agreement between MSAs. Dark yellow, orange, and red residues can be considered to be reliably aligned.



doi:10.1371/journal.pcbi.0030123.g002

Figure 2. Framework of a Template-Based Method

Structural templates are first identified, mapped onto the sequences, and aligned using SAP. The sequence–template mapping is then used to guide the alignment of the original sequences. This alignment is integrated into the library that is used to compute the final MSA.

RNA secondary structure predictions as templates and fills a T-Coffee library with the Lara pairwise algorithm. With the exception of PRALINE and SPEM, which use a regular progressive algorithm, most template-based methods described here are consistency-based (some of them taking advantage of T-Coffee modular structure). Their main advantage is increased accuracy. Recent benchmarks on PROMALS (Table 1) show that homology extension results in a ten-point improvement over existing methods. Likewise, structure-based methods such as Expresso produce alignments much closer to the structural references than do any of their sequence-based counterparts. One must, however, be careful not to over-interpret validation values like that given for Expresso in Table 1, since both the reference and the Expresso alignments were computed using the same structural information.

This last point raises the important issue of method validation and benchmarking. A recent study [4] shows that with the exception of artificial datasets, benchmarks carried out on most reference databases tend to deliver compatible results. It also suggests that the best methods have become indistinguishable, except when considering remote homologs (less than 25% identity). Unfortunately, remote homologs are poorly suited to generating reference alignment, owing to the fact that their superposition often yields alternative sequence alignments that are structurally equivalent [30]. However, one can bypass the reference alignment stage by directly comparing the evaluated alignment to some idealized 3-D superposition. Such an alignment-independent evaluation has been described and used by several authors [17,31,32]. Another trend, not well accounted for by current reference collections, is the alignment of very large datasets. While many new methods incorporate special algorithms for aligning several hundred sequences [6,8,18], current reference databases do not allow the evaluation of very large datasets, thus making it unclear how the published accuracies scale with the number of sequences. While this last issue could probably be satisfyingly solved in the current benchmarking framework, another problem remains that is much harder to address. All the existing validation approaches have in common their reliance on the "one size fits all" assumption that structurally correct alignments are the best possible MSAs for modeling any kind of biological signal (evolution, homology, or function). A report on profile construction [33] has recently challenged this view by showing that structurally correct alignments do not necessarily result in better profiles. Likewise, it may be reasonable to ask whether better alignments always result in better phylogenetic trees, and, more systematically, to question and quantify the relationship between the accuracy of MSAs and the biological relevance of any model drawn upon them.

In this review, I have presented some of the latest additions to the MSA computation arsenal. An interesting milestone has been the development of meta-methods able to seamlessly combine the output of several methods. Aside from easing the user's work, the main advantage of these consensus methods is probably the local estimation of reliability they provide (Figure 1). Using this estimation to filter out unreliable regions has already proven useful in homology modeling [34] and could probably be used further. The main improvement reported here, however, is probably the notion of template-based alignment. Template-based alignment is more than a trivial extension of consistency-based methods. Under this new model, the purpose of an MSA is not to squeeze a dataset and extract all the information it may contain, but rather to use the dataset as a starting point for exploring and retrieving all the related information contained in public databases. This information is to be used not only for mapping purposes, but also for driving the MSA computation. Such a usage of sequence information makes template-based methods a real paradigm shift and a major step toward global biological data integration. ■

Acknowledgments

The author thanks the two anonymous reviewers for suggesting several missing references.

Author contributions. CN analyzed the data and wrote the paper.

Funding. CN is funded and supported by the Centre National de la Recherche Scientifique, France.

Competing interests. The author has declared that no competing interests exist.

References

1. Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16: 368–373.
2. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Curr Opin Struct Biol 15: 261–266.

3. Gotoh O (1999) Multiple sequence alignment: Algorithms and applications. Adv Biophys 36: 159–206.
4. Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6: 321–339.
5. Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. J Mol Evol 20: 175–186.
6. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
7. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
8. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–518.
9. Simossis VA, Heringa J (2005) PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33: W289–W294.
10. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.
11. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385–395.
12. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34: 1692–1699.
13. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, et al. (2006) Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34: W604–W608.
14. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4690.
15. Pei J, Sadreyev R, Grishin NV (2003) PCMA: Fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19: 427–428.
16. Pei J, Grishin NV (2007) PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23: 802–808.
17. Pei J, Grishin NV (2006) MUMMALS: Multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 34: 4364–4374.
18. Lassmann T, Sonnhammer EL (2005) Kalign—An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298.
19. Lassmann T, Sonnhammer EL (2006) Kalign, Kalignvu and Mumsa: Web servers for multiple sequence alignment. Nucleic Acids Res 34: W596–W599.
20. Morgenstern B, Dress A, Wener T (1996) Multiple DNA and protein sequence based on segment-to-segment comparison. Proc Natl Acad Sci U S A 93: 12098–12103.
21. Notredame C, Abergel C (2003) Using multiple alignment methods to assess the quality of genomic data analysis. In: Andrade M, editor. Bioinformatics and genomes: Current perspectives. Wymondham (United Kingdom): Horizon Scientific Press. pp. 30–50.
22. Lassmann T, Sonnhammer EL (2005) Automatic assessment of alignment quality. Nucleic Acids Res 33: 7120–7128.
23. Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188: 233–258.
24. Thompson JD, Plewniak F, Thierry J, Poch O (2000) DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 28: 2919–2926.
25. Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208: 1–22.
26. Taylor WR (1999) Protein structure comparison using iterated double dynamic programming. Protein Sci 8: 654–665.
27. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243–257.
28. Zhou H, Zhou Y (2005) SPEM: Improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21: 3615–3621.
29. Bauer M, Klau G, Reinert K (2005) Multiple structural RNA alignment with Lagrangian relaxation. Lect Notes Comput Sci 3692: 303–314.
30. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS (2000) ProSup: A refined tool for protein structure alignment. Protein Eng 13: 745–752.
31. O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, et al. (2003) APDB: A novel measure for benchmarking sequence alignment methods without reference alignments. Bioinformatics 19: I215–I221.
32. Armougom F, Moretti S, Keduas V, Notredame C (2006) The iRMSD: A local measure of sequence alignment accuracy using structural information. Bioinformatics 22: e35–e39.
33. Griffiths-Jones S, Bateman A (2002) The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics 18: 1243–1249.
34. Claude JB, Suhre K, Notredame C, Claverie JM, Abergel C (2004) CaspR: A web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32: W606–W609.



D.2 Databases

Further explanations for Chapter 2.

D.2.1 How biological databases work

- Oza et al. (2023) — Practical advice on finding, judging, downloading and citing data from public databases, including metadata, versions and access restrictions. A short, broad introduction to the issues discussed in this section. *(tutorial; CC BY 4.0)*

  Ten simple rules for using public biological data for your research. Oza et al. (2023). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1010749

  > With an increasing amount of biological data available publicly, there is a need for a guide on how to successfully download and use this data. The 10 simple rules for using public biological data are: (1) use public data purposefully in your research; (2) evaluate data for your use case; (3) check data reuse requirements and embargoes; (4) be aware of ethics for data reuse; (5) plan for data storage and compute requirements; (6) know what you are downloading; (7) download programmatically and verify integrity; (8) properly cite data; (9) make reprocessed data and models Findable, Accessible, Interoperable, and Reusable (FAIR) and share; and (10) make pipelines and code FAIR and share. These rules are intended as a guide for researchers wanting to make use of available data and to increase data reuse and reproducibility.

- Danielewski et al. (2025) — Reviews why and how biological sequence databases were established, the move to open access, and current policy questions such as privacy and the international debate on digital sequence information. Useful background for the subsection on licences and access. *(review; CC BY 4.0)*

  History of Biological Databases, Their Importance, and Existence in Modern Scientific and Policy Context. Danielewski et al. (2025). *Genes*. doi:10.3390/genes16010100

  > With the development of genome sequencing technologies, the amount of data produced has greatly increased in the last two decades. The abundance of digital sequence information (DSI) has provided research opportunities, improved our understanding of the genome, and led to the discovery of new solutions in industry and medicine. It has also posed certain challenges, i.e., how to store and handle such amounts of data. This, coupled with the need for convenience, international cooperation, and the possibility of independent validation, has led to the establishment of numerous databases. Spearheaded with the idea that data obtained with public funds should be available to the public, open access has become the predominant mode of accession. However, the increasing popularity of commercial genetic tests brings back the topic of data misuse, and patient's privacy. At the previous United Nations Biodiversity Conference (COP15, 2022), an issue of the least-developed countries exploiting their natural resources while providing DSI and the most-developed countries benefitting from this was raised. It has been proposed that financial renumeration for the data could help protect biodiversity. With the goal of introducing the topic to those interested in utilizing biological databases, in this publication, we present the history behind the biological databases, their necessity in today's scientific world, and the issues that concern them and their content, while providing scientific and policy context in relation to United Nations Biodiversity Conference (COP16, 21.10-1.11.24).

- Rosonovski et al. (2023) — Step-by-step protocols for searching the literature, full texts and preprints in Europe PMC, including programmatic access. Complements the subsection on PubMed, PMC and preprints. *(tutorial; CC BY 4.0)*

  Searching and Evaluating Publications and Preprints Using Europe PMC. Rosonovski et al. (2023). *Current Protocols*. doi:10.1002/cpz1.694

  > In the field of life sciences there is a growing need for literature analysis tools that help scientists tackle information overload. Europe PubMed Central (Europe PMC), a partner of PubMed Central (PMC; National Library of Medicine, 2022), is an open access database of over 41 million life science publications and preprints, enriched with supporting data, reviews, protocols, and other relevant resources. Europe PMC is a trusted repository of choice for many life science funders (Europe PMC, 2022a), offering a suite

of innovative search tools that allow users to search and evaluate the literature, including finding highly cited articles, preprints with community peer reviews, or papers referencing a proteomics dataset in the figure legend. In addition, Europe PMC utilizes text-mining to help researchers identify key terms and find data and evidence in the literature. First-time users often do not utilize the wealth of tools Europe PMC offers and can feel overwhelmed about how to perform the most effective search. This protocol, describing how to search and evaluate publications and preprints using Europe PMC, demonstrates how to carry out more efficient and effective literature searches using the tools provided by Europe PMC. This includes discovering the latest findings on a research topic, following research from a specific author, journal, or preprint server, exploring literature on a new method, expanding your reading list with relevant articles, as well as accessing and evaluating publications and preprints of interest. © 2023 EMBL-EBI. Current Protocols published by Wiley Periodicals LLC. Basic Protocol 1: Finding articles and preprints on a topic of interest Basic Protocol 2: Accessing an article Basic Protocol 3: Browsing the article Basic Protocol 4: Evaluating the article Basic Protocol 5: Refining search results Basic Protocol 6: Finding research by author Basic Protocol 7: Finding a specific article Basic Protocol 8: Finding information about a methodology Basic Protocol 9: Finding evidence of biological interactions, relations, and modifications Basic Protocol 10: Finding data behind a publication Basic Protocol 11: Expanding a reading list and building a bibliography Basic Protocol 12: Staying on top of the current literature.

D.2.2 Sequence databases

- Lussi et al. (2023) — Protocols written by the UniProt team on navigating and searching the UniProt datasets (UniProtKB, UniRef, UniParc and Proteomes) and retrieving data. A practical complement to the description of UniProt in this section. *(tutorial; CC BY 4.0)*

  Searching and Navigating UniProt Databases. Lussi et al. (2023). *Current Protocols*. doi:10.1002/cpz1.700

  > The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt website receives about 800,000 unique visitors per month and is the primary means to access UniProt. It provides 10 searchable datasets and four main tools. The key UniProt datasets are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), and protein sets for completely sequenced genomes (Proteomes). Other supporting datasets include information about proteins that is present in UniProtKB protein entries, such as literature citations, taxonomy, and subcellular locations, among others. This article focuses on how to use UniProt datasets. The first basic protocol describes navigation and searching mechanisms for the UniProt datasets, and two additional protocols build on the first protocol to describe advanced search and query building. © 2023 The Authors. Current Protocols published by Wiley Periodicals LLC. Basic Protocol 1: Searching UniProt datasets Basic Protocol 2: Advanced search and query building Basis Protocol 3: Adding parameters using advanced search.

- Vuong et al. (2022) — Guidance on working with environmental metagenomic data, including sample metadata, reference databases and data deposition. Relevant to the subsection on metagenomic sequences. *(tutorial; CC BY 4.0)*

  Ten simple rules for investigating (meta)genomic data from environmental ecosystems. Vuong et al. (2022). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1010675

  > Metagenomics is the sequencing and study of DNA directly from environmental samples and is an approach used in microbial ecology to explore the diversity and functions of microbial communities [1].The study of the metagenome presents an avenue to investigate environmental microorganisms in-situ through culture-independent techniques, which facilitates the discovery and analysis of the metabolic potential of hitherto uncultured taxa [2].Expanding our reach into the once obscure facets of microbial life has yielded further insights into the ecological activity of microorganisms, uncovering a trove of novel metabolic products and pathways with a wide range of biotechnical applications in many industries [3,4].Data obtained from metagenomes have also been pivotal in directing strategies for the cultivation of novel microorganisms [5].The benefit of metagenomics is that it forms a wide-reaching preliminary approach that can provide contextual information to direct further downstream research for microorganisms in the wet lab.Microbial ecology has developed to a stage where ”big data” is often the norm.Through advances in sequencing and computational technologies, we have experienced an unprecedented level of progress in the field of microbial ecology, improving our understanding of the many global ecosystems [6].Metagenomics draws many parallels when compared to data science-environmental sequence data

> presents a large volume of information that must be sifted through and sorted to make sense of underlying ecological and taxonomic patterns [7].The amount of information that needs to be processed, however, can be daunting for those starting out and often still pose challenges even for those familiar to the field.The inspiration behind this writing is from a recommendation to a PLOS Ten Simple Rules article on command-line bioinformatics [8].The article provided a lot of information on tools and concepts used by bioinformaticians and it would have been a valuable source of guidance for many.This paper styled in the typical format of a Ten Simple Rules article, presents a guide on the basics of metagenomic-based research aimed at environmental settings.We will explore the concepts, tools, and resources utilized in metagenomics and how these can potentially help the progress of a researcher beginning their studies (Fig 1).The 10 simple rules format is a casual one and people are often busy, so at the end of each rule is a TL; DR (too long; didn't read) that summarizes the point we are trying to make if you want to skim through (although we do hope you'll stick around and read the paper in depth!).

D.2.3 Genome databases and browsers

- Spudich and Fernández-Suárez (2010) — A worked tour of the Ensembl browser through four examples covering gene pages, comparative genomics, disease-linked variation and functional genomics. The interface has changed since 2010, but the concepts it explains still apply. *(tutorial; CC BY (Europe PMC))*

  Touring Ensembl: A practical guide to genome browsing. Spudich & Fernández-Suárez (2010). *BMC Genomics*. doi:10.1186/1471-2164-11-295

  > The number of databases in molecular biological fields has rapidly increased to provide a large-scale resource. Though valuable information is available, data can be difficult to access, compare and integrate due to different formats and presentations of web interfaces. This paper offers a practical guide to the integration of gene, comparative genomic, and functional genomics data using the Ensembl website at http://www.ensembl.org.The Ensembl genome browser and underlying databases focus on chordate organisms. More species such as plants and microorganisms can be investigated using our sister browser at http://www.ensemblgenomes.org.In this study, four examples are used that sample many pages and features of the Ensembl browser. We focus on comparative studies across over 50 mostly chordate organisms, variations linked to disease, functional genomics, and access of external information housed in databases outside the Ensembl project. Researchers will learn how to go beyond simply exporting one gene sequence, and explore how a genome browser can integrate data from various sources and databases to build a full and comprehensive biological picture.

- Oliver et al. (2016) — Explains what model organism databases do, why expert curation of the literature matters, and the problem of funding them. Background for the subsection on organism-specific databases. *(perspective; CC BY (Europe PMC))*

  Model organism databases: essential resources that need the support of both funders and users. Oliver et al. (2016). *BMC Biology*. doi:10.1186/s12915-016-0276-z

  > Modern biomedical research depends critically on access to databases that house and disseminate genetic, genomic, molecular, and cell biological knowledge. Even as the explosion of available genome sequences and associated genome-scale data continues apace, the sustainability of professionally maintained biological databases is under threat due to policy changes by major funding agencies. Here, we focus on model organism databases to demonstrate the myriad ways in which biological databases not only act as repositories but actively facilitate advances in research. We present data that show that reducing financial support to model organism databases could prove to be not just scientifically, but also economically, unsound.

- Dominguez Del Angel et al. (2018) — An overview of how a genome is assembled and annotated before it appears in genome databases. Helps to understand where gene models and assembly versions come from. *(tutorial; CC BY 4.0)*

  Ten steps to get started in Genome Assembly and Annotation. Dominguez Del Angel et al. (2018). *F1000Research*. doi:10.12688/f1000research.13598.1

> As a part of the ELIXIR-EXCELERATE efforts in capacity building, we present here 10 steps to facilitate researchers getting started in genome assembly and genome annotation. The guidelines given are broadly applicable, intended to be stable over time, and cover all aspects from start to finish of a general assembly and annotation project. Intrinsic properties of genomes are discussed, as is the importance of using high quality DNA. Different sequencing technologies and generally applicable workflows for genome assembly are also detailed. We cover structural and functional annotation and encourage readers to also annotate transposable elements, something that is often omitted from annotation workflows. The importance of data management is stressed, and we give advice on where to submit data and how to make your results Findable, Accessible, Interoperable, and Reusable (FAIR).

D.2.4 Structure databases

- David et al. (2022) — Explains how to use and interpret predicted structures from the AlphaFold Database, including the pLDDT and PAE confidence measures. Useful for the subsection on model databases. *(review; CC BY (Europe PMC))*

  The AlphaFold Database of Protein Structures: A Biologist's Guide. David et al. (2022). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2021.167336

  > AlphaFold, the deep learning algorithm developed by DeepMind, recently released the three-dimensional models of the whole human proteome to the scientific community. Here we discuss the advantages, limitations and the still unsolved challenges of the AlphaFold models from the perspective of a biologist, who may not be an expert in structural biology.

- Burley et al. (2024) — Describes the PDB archive, its deposition and validation, and the RCSB tools for searching and visualising both experimental structures and computed models. *(review; CC BY 4.0)*

  RCSB Protein Data Bank: supporting research and education worldwide through explorations of experimentally determined and computationally predicted atomic level 3D biostructures. Burley et al. (2024). *IUCrJ*. doi:10.1107/s2052252524002604

  > The Protein Data Bank (PDB) was established as the first open-access digital data resource in biology and medicine in 1971 with seven X-ray crystal structures of proteins. Today, the PDB houses >210 000 experimentally determined, atomic level, 3D structures of proteins and nucleic acids as well as their complexes with one another and small molecules (e.g. approved drugs, enzyme cofactors). These data provide insights into fundamental biology, biomedicine, bioenergy and biotechnology. They proved particularly important for understanding the SARS-CoV-2 global pandemic. The US-funded Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) and other members of the Worldwide Protein Data Bank (wwPDB) partnership jointly manage the PDB archive and support >60 000 'data depositors' (structural biologists) around the world. wwPDB ensures the quality and integrity of the data in the ever-expanding PDB archive and supports global open access without limitations on data usage. The RCSB PDB research-focused web portal at https://www.rcsb.org/ (RCSB.org) supports millions of users worldwide, representing a broad range of expertise and interests. In addition to retrieving 3D structure data, PDB 'data consumers' access comparative data and external annotations, such as information about disease-causing point mutations and genetic variations. RCSB.org also provides access to >1 000 000 computed structure models (CSMs) generated using artificial intelligence/machine-learning methods. To avoid doubt, the provenance and reliability of experimentally determined PDB structures and CSMs are identified. Related training materials are available to support users in their RCSB.org explorations.

- Varadi et al. (2022) — Describes the European PDB node and its knowledge base, which links structures to sequence, function and ligand annotation. Shows how structure data are integrated with other databases. *(review; CC BY 4.0)*

  PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. Varadi et al. (2022). *Protein Science*. doi:10.1002/pro.4439

> The archiving and dissemination of protein and nucleic acid structures as well as their structural, functional and biophysical annotations is an essential task that enables the broader scientific community to conduct impactful research in multiple fields of the life sciences. The Protein Data Bank in Europe (PDBe; pdbe.org) team develops and maintains several databases and web services to address this fundamental need. From data archiving as a member of the Worldwide PDB consortium (wwPDB; wwpdb.org), to the PDBe Knowledge Base (PDBe-KB; pdbekb.org), we provide data, data-access mechanisms, and visualizations that facilitate basic and applied research and education across the life sciences. Here, we provide an overview of the structural data and annotations that we integrate and make freely available. We describe the web services and data visualization tools we offer, and provide information on how to effectively use or even further develop them. Finally, we discuss the direction of our data services, and how we aim to tackle new challenges that arise from the recent, unprecedented advances in the field of structure determination and protein structure modeling.

D.2.5 Function and pathways

- Gaudet et al. (2016) — A short introduction to the structure of GO, its three aspects, relations between terms and how annotations with evidence codes are made. A good companion to the Gene Ontology subsection. *(book chapter; CC BY 4.0)*

  Primer on the Gene Ontology. Gaudet et al. (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_3

  > The Gene Ontology (GO) project is the largest resource for cataloguing gene function. The combination of solid conceptual underpinnings and a practical set of features have made the GO a widely adopted resource in the research community and an essential resource for data analysis. In this chapter, we provide a concise primer for all users of the GO. We briefly introduce the structure of the ontology and explain how to interpret annotations associated with the GO.

- Rothfels et al. (2023) — Protocols from the Reactome team on browsing pathways, interpreting the pathway diagrams and running pathway analysis on a gene list. *(tutorial; CC BY 4.0)*

  Using the Reactome Database. Rothfels et al. (2023). *Current Protocols*. doi:10.1002/cpz1.722

  > Pathway databases provide descriptions of the roles of proteins, nucleic acids, lipids, carbohydrates, and other molecular entities within their biological cellular contexts. Pathway-centric views of these roles may allow for the discovery of unexpected functional relationships in data such as gene expression profiles and somatic mutation catalogues from tumor cells. For this reason, there is a high demand for high-quality pathway databases and their associated tools. The Reactome project (a collaboration between the Ontario Institute for Cancer Research, New York University Langone Health, the European Bioinformatics Institute, and Oregon Health & Science University) is one such pathway database. Reactome collects detailed information on biological pathways and processes in humans from the primary literature. Reactome content is manually curated, expert-authored, and peer-reviewed and spans the gamut from simple intermediate metabolism to signaling pathways and complex cellular events. This information is supplemented with likely orthologous molecular reactions in mouse, rat, zebrafish, worm, and other model organisms. © 2023 The Authors. Current Protocols published by Wiley Periodicals LLC. Basic Protocol 1: Browsing a Reactome pathway Basic Protocol 2: Exploring Reactome annotations of disease and drugs Basic Protocol 3: Finding the pathways involving a gene or protein Alternate Protocol 1: Finding the pathways involving a gene or protein using UniProtKB (SwissProt), Ensembl, or Entrez gene identifier Alternate Protocol 2: Using advanced search Basic Protocol 4: Using the Reactome pathway analysis tool to identify statistically overrepresented pathways Basic Protocol 5: Using the Reactome pathway analysis tool to overlay expression data onto Reactome pathway diagrams Basic Protocol 6: Comparing inferred model organism and human pathways using the Species Comparison tool Basic Protocol 7: Comparing tissue-specific expression using the Tissue Distribution tool.

- Chicco and Agapito (2022) — Practical advice on using GO and pathway databases for enrichment analysis, including choice of background and database version. Shows common pitfalls when function databases are used for interpretation. *(tutorial; CC BY 4.0)*

  Nine quick tips for pathway enrichment analysis. Chicco & Agapito (2022). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1010348

Pathway enrichment analysis (PEA) is a computational biology method that identifies biological functions that are overrepresented in a group of genes more than would be expected by chance and ranks these functions by relevance. The relative abundance of genes pertinent to specific pathways is measured through statistical methods, and associated functional pathways are retrieved from online bioinformatics databases. In the last decade, along with the spread of the internet, higher availability of computational resources made PEA software tools easy to access and to use for bioinformatics practitioners worldwide. Although it became easier to use these tools, it also became easier to make mistakes that could generate inflated or misleading results, especially for beginners and inexperienced computational biologists. With this article, we propose nine quick tips to avoid common mistakes and to out a complete, sound, thorough PEA, which can produce relevant and robust results. We describe our nine guidelines in a simple way, so that they can be understood and used by anyone, including students and beginners. Some tips explain what to do before starting a PEA, others are suggestions of how to correctly generate meaningful results, and some final guidelines indicate some useful steps to properly interpret PEA results. Our nine tips can help users perform better pathway enrichment analyses and eventually contribute to a better understanding of current biology.

D.2.6 Interaction databases

- De Las Rivas and Fontanillo (2010) — Explains the types of protein interaction data, the experimental methods behind them and the primary and integrated interaction databases. Older, but clear on the distinction between physical and functional interactions. *(tutorial; CC BY 4.0)*

  Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. De Las Rivas & Fontanillo (2010). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000807

- Panneerselvam et al. (2024) — Protocols for searching IntAct and using the detailed curation of experimental context that IMEx databases record. Illustrates what a curated physical interaction record contains. *(tutorial; CC BY 4.0)*

  IntAct Database for Accessing IMEx's Contextual Metadata of Molecular Interactions. Panneerselvam et al. (2024). *Current Protocols*. doi:10.1002/cpz1.70018

  The International Molecular Exchange Consortium (IMEx) has evolved into a vital partnership of open resources dedicated to curating molecular interaction data from the scientific literature. This consortium, which includes IntAct, MINT, MatrixDB, and DIP, is a collaborative effort with a central mission of aggregating detailed molecular interaction experimental evidence in a machine-readable format, supported by controlled vocabularies and standard ontologies. The IntAct molecular interaction database (www.ebi.ac.uk/intact), as an IMEx partner, serves as a valuable portal for accessing IMEx data through user-friendly search options and an array of interactive filters. The resource currently hosts an extensive repository of 1,293,508 binary interactions meticulously captured from 75,098 experiments documented in 23,366 publications (as of the February 2024 release), with this corpora being added to by regular data releases. IMEx curation policy has consistently prioritized a fine-grained data and curation model, with a focus on capturing the relevant experimental details essential for interpreting molecular interaction data effectively. Our curation process is designed to support the generation of interactomes tailored to contexts such as disease-specific or tissue-/cell-type-specific interactomes. These interactions are ranked according to a scoring system based on the Proteomics Standard Initiative Molecular Interaction (PSI MI) standards. This scoring system allows users to assess the degree of confidence in binary interactions, enhancing the value of the data. The resource provides insights into the nature of relationships among interacting partners as defined by the experimental setup and the associated biological context. Interactive filters enable users to navigate these rich, multilayered data, promoting a deeper understanding of biological complexity. Additionally, the IntAct website fosters the creation of networks for collaborative analyses by the scientific community. The recent transformation of the IntAct website, supported by a graph-type database, empowers users to execute custom queries tailored to their specific research interests. This article illustrates the diverse levels of annotations available for interactions and the multiple search options at users' disposal to access data of interest. © 2024 European Molecular Biology Laboratory, European Bioinformatics Institute. Current Protocols published by Wiley Periodicals LLC. Basic Protocol 1: Using Quick Search, network visualization, and filters Support Protocol: Accessing fine annotations from intact: Unlocking the molecular details Alternate Protocol: Using batch search: Querying multiple interactors Basic Protocol 2: Using advanced search: Precision and customization.

D.2.7 Expression and proteomics repositories

- Wang et al. (2018) — Reviews methods for curating and reprocessing GEO datasets at scale, and explains how poorly standardised metadata limits reuse. Useful for understanding both the value and the limitations of expression repositories. *(review; CC BY 4.0)*

  Mining data and metadata from the gene expression omnibus. Wang et al. (2018). *Biophysical Reviews*. doi:10.1007/s12551-018-0490-8

  > Publicly available gene expression datasets deposited in the Gene Expression Omnibus (GEO) are growing at an accelerating rate. Such datasets hold great value for knowledge discovery, particularly when integrated. Although numerous software platforms and tools have been developed to enable reanalysis and integration of individual, or groups, of GEO datasets, large-scale reuse of those datasets is impeded by minimal requirements for standardized metadata both at the study and sample levels as well as uniform processing of the data across studies. Here, we review methodologies developed to facilitate the systematic curation and processing of publicly available gene expression datasets from GEO. We identify trends for advanced metadata curation and summarize approaches for reprocessing the data within the entire GEO repository.

- Perez-Riverol et al. (2015) — An overview of proteomics repositories, ProteomeXchange and data standards. Some details are dated, but it explains the landscape that PRIDE and related resources belong to. *(review; CC BY 4.0)*

  Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Perez-Riverol et al. (2015). *PROTEOMICS*. doi:10.1002/pmic.201400302

  > Compared to other data-intensive disciplines such as genomics, public deposition and storage of MS-based proteomics, data are still less developed due to, among other reasons, the inherent complexity of the data and the variety of data types and experimental workflows. In order to address this need, several public repositories for MS proteomics experiments have been developed, each with different purposes in mind. The most established resources are the Global Proteome Machine Database (GPMDB), PeptideAtlas, and the PRIDE database. Additionally, there are other useful (in many cases recently developed) resources such as ProteomicsDB, Mass Spectrometry Interactive Virtual Environment (MassIVE), Chorus, MaxQB, PeptideAtlas SRM Experiment Library (PASSEL), Model Organism Protein Expression Database (MOPED), and the Human Proteinpedia. In addition, the ProteomeXchange consortium has been recently developed to enable better integration of public repositories and the coordinated sharing of proteomics information, maximizing its benefit to the scientific community. Here, we will review each of the major proteomics resources independently and some tools that enable the integration, mining and reuse of the data. We will also discuss some of the major challenges and current pitfalls in the integration and sharing of the data.

D.2.8 Medical and variant data

- Riccio et al. (2024) — A systematic review of 118 databases and software tools for annotating the functional impact of genetic variants, with practical advice on which to combine. Relevant to the subsection on variants and their interpretation. *(review; CC BY 4.0)*

  Variant effect predictors: a systematic review and practical guide. Riccio et al. (2024). *Human Genetics*. doi:10.1007/s00439-024-02670-5

  > Large-scale association analyses using whole-genome sequence data have become feasible, but understanding the functional impacts of these associations remains challenging. Although many tools are available to predict the functional impacts of genetic variants, it is unclear which tool should be used in practice. This work provides a practical guide to assist in selecting appropriate tools for variant annotation. We conducted a MEDLINE search up to November 10, 2023, and included tools that are applicable to a broad range of phenotypes, can be used locally, and have been recently updated. Tools were categorized based on the types of variants they accept and the functional impacts they predict. Sequence Ontology terms were used for standardization. We identified 118 databases and software packages, encompassing 36 variant types and 161 functional impacts. Combining only three tools, namely SnpEff, FAVOR, and SparkINFERNO, allows predicting 99 (61%) distinct functional impacts. Thirty-seven tools predict 89 functional impacts that are not supported by any other tool, while 75 tools predict pathogenicity and

> can be used within the ACMG/AMP guidelines in a clinical context. We launched a website allowing researchers to select tools based on desired variants and impacts. In summary, more than 100 tools are already available to predict approximately 160 functional impacts. About 60% of the functional impacts can be predicted by the combination of three tools. Unexpectedly, recent tools do not predict more impacts than older ones. Future research should allow predicting the functionality of so far unsupported variant types, such as gene fusions.URL: https://cardio-care.shinyapps.io/VEP_Finder/ .Registration: OSF Registries on November 10, 2023, https://osf.io/s2gct .

- Corpas et al. (2018) — Guidelines for providers of human genomic data on making shared datasets findable, accessible, interoperable and reusable. Background for the subsection on controlled access to personal data. *(tutorial; CC BY 4.0)*

  A FAIR guide for data providers to maximise sharing of human genomic data. Corpas et al. (2018). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1005873

  > It is generally acknowledged that, for reproducibility and progress of human genomic research, data sharing is critical. For every sharing transaction, a successful data exchange is produced between a data consumer and a data provider. Providers of human genomic data (e.g., publicly or privately funded repositories and data archives) fulfil their social contract with data donors when their shareable data conforms to FAIR (findable, accessible, interoperable, reusable) principles. Based on our experiences via Repositive (https://repositive.io), a leading discovery platform cataloguing all shared human genomic datasets, we propose guidelines for data providers wishing to maximise their shared data's FAIRness.

- Van Buren et al. (2025) — Compares human short-variant databases with those built for animal species, covering methods, data integration and applications. Gives a wider view of how variant repositories are organised. *(review; CC BY-NC 4.0 (link only))*

  A comparative review of short genetic variant databases across humans and animal species. Van Buren et al. (2025). *Briefings in Bioinformatics*. doi:10.1093/bib/bbaf356

  > Understanding genetic variation is necessary to unravel the complexities of evolution and diverse traits across species. Short genetic variants (<50 bp in length) represent key genomic variations that play a crucial role in shaping the genetic landscape of populations. Human short genetic variant databases are often considered the gold standard for variant repositories, and this review compares them with variant databases created for various animal species. This review examines the methodologies, data integration, and various applications that differ or align between human and animal databases. The goal is to identify challenges in leveraging genetic variation across species, recommend strategies for overcoming these challenges, and suggest future directions for research and database development.

The following open-access paper(s) are reproduced under their original licences.

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Ten simple rules for using public biological data for your research

Vishal H. Oza[1], Jordan H. Whitlock[1], Elizabeth J. Wilk[1], Angelina Uno-Antonison[2,3,4], Brandon Wilk[2,3,4], Manavalan Gajapathy[2,3,4], Timothy C. Howton[1], Austyn Trull[2,3,4], Lara Ianov[5], Elizabeth A. Worthey[2,3,4], Brittany N. Lasseigne[1]*

1 Department of Cell, Developmental and Integrative Biology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **2** Center for Computational Genomics and Data Sciences, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **3** Department of Pediatrics, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **4** Department of Pathology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States of America, **5** Civitan International Research Center, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States of America

* bnp0001@uab.edu

Check for updates

OPEN ACCESS

Citation: Oza VH, Whitlock JH, Wilk EJ, Uno-Antonison A, Wilk B, Gajapathy M, et al. (2023) Ten simple rules for using public biological data for your research. PLoS Comput Biol 19(1): e1010749. https://doi.org/10.1371/journal.pcbi.1010749

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Published: January 5, 2023

Copyright: © 2023 Oza et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: VHO, EJW, AUA, LI, and EAW, and BNL are supported by U54OD030167 (www.nih.gov). VHO, TCH, and BNL are supported by R00HG009678 (www.nih.gov). VHO and BNL are supported by R03OD030604 (www.nih.gov). JHW is supported by T32GM008111 (www.nih.gov). JHW, TCH, BNL are supported by UAB Lasseigne Lab Startup funds (www.uab.edu). AUO, BW, MG, AT, and EAW are supported by UAB Worthey Lab funds (www.uab.edu). BW and EAW are supported by UG1HD107688 (www.nih.gov). BW, MG, and EAW are supported by the CF Foundation (WORTHE19A0) (www.cff.org). LI is supported by UAB CIRC funds (www.uab.edu). AT, LI, EAW, and BNL are members of the UAB Biological Data Science (U-BDS) Core (RRID:SCR_021766) (www.

Abstract

With an increasing amount of biological data available publicly, there is a need for a guide on how to successfully download and use this data. The 10 simple rules for using public biological data are: (1) use public data purposefully in your research; (2) evaluate data for your use case; (3) check data reuse requirements and embargoes; (4) be aware of ethics for data reuse; (5) plan for data storage and compute requirements; (6) know what you are downloading; (7) download programmatically and verify integrity; (8) properly cite data; (9) make reprocessed data and models Findable, Accessible, Interoperable, and Reusable (FAIR) and share; and (10) make pipelines and code FAIR and share. These rules are intended as a guide for researchers wanting to make use of available data and to increase data reuse and reproducibility.

This is a *PLOS Computational Biology* Methods paper.

Introduction

In recent years, with the advent of high-throughput sequencing technologies, advances in microscopy, and the growth of single-cell technologies, biology is set to overtake other data-heavy disciplines such as astronomy in terms of data storage needs [1]. There has been a dramatic increase in the number and size of data sets deposited by individual labs on data storage servers like the Gene Expression Omnibus (GEO) and dbGAP [2,3] and made available by large consortium efforts such as The Cancer Genome Atlas (TCGA) [4], the Genotype-Tissue Expression (GTEx) project [5], Bgee [6], Human Cell Atlas [7], and ENCODE [8,9]. These resources provide rich and diverse data to the biology research community. For example,

uab.edu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

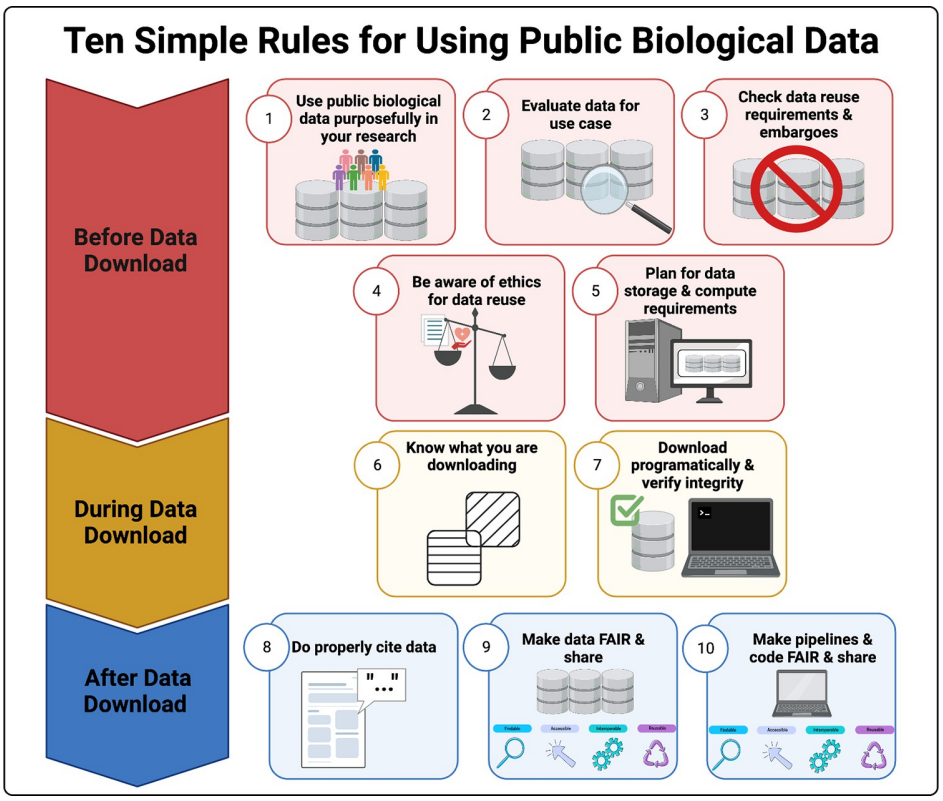


Fig 1. These 10 simple rules for using public data span checklist items for pre, during, and post data download.

https://doi.org/10.1371/journal.pcbi.1010749.g001

TCGA is made up of clinical information (e.g., smoking status), molecular analyte data (e.g., sample portion weight), and molecular characterization data (e.g., imaging data, omics data). The additional commitment by funders, publishers, and individual scientists to make data sets (especially those funded by taxpayers and donors) publicly available has rapidly increased opportunities for data reuse by the broader scientific community. From January 2023, NIH will now require researchers to share data generated with NIH funds under the new Data Management and Sharing Policy [10] further leading to an increase in the availability of public data. With this growing data deluge, it seemed timely to provide guidelines on why, when, and how investigators can incorporate these valuable resources into their research programs as biological data reuse is good for science, cost efficient, and is the right thing to do in order to extract the greatest societal impact from the samples and funding that patients, donors, and taxpayers generously provide. Here, we draw upon our collective experience and expertise as molecular biologists, computational biologists, bioinformaticians, data scientists, and software developers to discuss Ten Simple Rules for using public data with the intention that it will serve as a useful guide (Fig 1). While this article focuses on computational biology and bioinformatics, the principles outlined here generally apply to other domains as well.

Rule 1: Use public biological data purposefully in your research

Reasons for reusing public biological research data include cost-effectiveness and efficiency, access to data sets that would be difficult or impossible to regenerate, an increased sense of community, greater transparency and clarity of research, ability to retest and validate a shared

data set, support for recognition of data ownership, and an increase in citations [11]. Critically, public biological data reuse can also promote workforce diversity and research inclusivity worldwide [12–14]. Use of public biological data provides individuals with a unique opportunity to leverage existing data sets in combination with other public data or their own. This allows for additional contexts to be explored within your own research. Comparing your research findings with a public data set analyzed in the same way can give nuance or clarity to results, especially in either further confirming what was observed (i.e., validation) or producing different results that help you to re-evaluate your approach or its applicability across cohorts. For example, you can validate current findings in public data, expand the number of samples, age groups, or other parameters of your current analysis, or explore molecular changes in a different system. Assessing public data can also aid in placing your research into the context of, and in perspective to, current research with public data therefore allowing comparisons to other work in the field.

Incorporating public biological data in your research project can also allow you to ask novel questions that the data generator may not have originally foreseen, further moving your own research and the field in new directions. Outside of expanding your research through exploration of new contexts, public data sets can also be used to generate hypotheses and/or guide future research by refining hypotheses for preliminary analyses. However, be mindful of data being appropriate for answering your hypotheses as outlined in Rule 2. With computational biology becoming ubiquitous in life science research, you can also use public data to drive novel method development and build modeling frameworks for systems biology. In fact, there have been efforts to standardize data sets for ease of use in such analysis [15]. In addition to improving your analysis, using public data also provides an opportunity to practice reproducible and interoperable research and further develop professional data science skills [11]. However, it is important to avoid the "sins of methodological research" (i.e., selective reporting, lack of replication studies, poor design of comparison studies, publication bias) [16] and therefore we suggest ensuring your research team has the proper methodological expertise to properly use and interpret public biological data sets.

In short, novel studies do not always require new data, and research including data reuse not only offers all of the benefits mentioned here, but can support community building and greater impact; as shown by Milham and colleagues [17] who found that neuroimaging research making use of public data was cited at the same or higher frequency in higher impact journals as compared to research using only self-generated data (this is expanded upon more in Rule 8). By using public biological data with care, your research can benefit from these benefits.

Rule 2: Evaluate data carefully for your use case

Before we get down to the business of downloading the data, we need to make sure we access appropriate data. As scientists, we are interested in finding signals and/or patterns in our data. However, patterns in a data set can arise from many different variables. Therefore, it is essential to carefully evaluate the available data for your particular use case; otherwise, the results obtained could be wrong and/or misleading. It is critical to examine the metadata (the data describing the data set) before downloading the actual data. Some important metadata variables to consider (but by no means exhaustive) relate to the (1) nature of the sample (e.g., the specific animal model or cell line(s), genotype, age, sex); (2) sample collection data (e.g., the type of sample collected, time points of collection); (3) platforms used (e.g., the sequencing or imaging platform); and (4) data quality metrics (e.g., sample size, groups of comparison). Not considering these variables can dramatically influence downstream analysis and interpretation

and should therefore be considered before choosing the appropriate data set to use. For more details, see Rules 4 and 5 in "Ten simple rules for providing effective bioinformatics research support" [18].

Once the data set is deemed appropriate for your analytical goals, it is key to check for the confounding effects of metadata variables that might impact the statistical model and cause you to reach wrong conclusions. This includes review of data distributions and sampling as well as performing exploratory analysis and visualizations such as principal component analysis (PCA) and correlation analysis. "Ten simple rules for initial data analysis"[19] provides excellent pointers for performing initial data screening including how to craft a reproducible plan and pitfalls to avoid.

Rule 3: Be aware of, and adhere to, data reuse requirements, embargoes, etc.

Researchers looking to reuse controlled-access data (elaborated more in Rule 4) assume the responsibility (along with their legal and financial administrative offices) to protect the rights and welfare of the original participants. As such, data repositories with controlled-access data sets develop Data Access Request (DAR) processes to appropriately restrict access. DAR processes usually include agreeing to terms of access for the requested data, a Data Use Certification Agreement (DUA) between the requester's institution and the repository, a description of the intended use, and an acknowledgment agreement. If a data set is available in a data portal (e.g., GEO [2], SRA [20]) that does not require controlled access, it is considered open access and can be freely downloaded through the portal. Even under these circumstances, the original data generators may still request acknowledgment upon use (a practice that should be followed whether or not the generators have requested it). A third possibility for acquiring controlled-access data for reuse is to contact the corresponding author of the paper directly to request data. However, this method can prove difficult for many reasons related to the author's availability to fulfill the request (e.g., protected time to fulfill the request, ability to transfer the data, willingness to comply). Therefore, independent of the method of requesting access to restricted data, it is important to be patient, but persistent. Recently, many journals have begun implementing data availability manuscript sections that outline the portal where the data is stored and can be accessed.

Regardless of how the data is acquired, it is important to be aware of the legal, regulatory, and security obligations associated with its use. For example, data licenses might be in place to protect the original data generator's rights by permitting secondary parties to reuse the data according to specific restrictions. Because countries have different regulations for data reuse [21], data licenses may clarify the uncertainty of requirements for data reusers. It is important to understand by which license or waiver the data to be reused is governed. In an effort to promote open access to data, many journals and data repositories operate under Creative Commons (CC BY 4.0) licenses that allow the freedom to share and adapt data as long as the original data generators are acknowledged for their contribution. For example, PLOS ONE stipulates that if the data associated with a published article in their journal is deposited in a repository with a licensing agreement, the agreement cannot be more restrictive than CC BY [22]. Researchers should identify which, if any, data license is governing the data they wish to reuse and respect any limitations associated with it.

Some data generators place an embargo on the data they generate in order to ensure they have time to publish initial findings. The data set may be submitted to a public repository but unavailable for download or publication for a certain length of time or until a specified date. Additionally, with the increased prevalence of pre-printing articles, data generators may wish

to withhold their data sets until their article is published in a peer-reviewed journal. Being aware of data access and publication restrictions associated with the data set of interest and understanding who sets data restrictions can be complicated (e.g., funders, consortia, journals, individual labs), but convention, or in some cases the legal requirement, is to follow the stated restrictions.

Rule 4: Be aware of ethical considerations like confidentiality and protected health information

The privacy and ethics of data sharing are critical and it is the responsibility of researchers to protect and observe [23]. With regards to data reuse, you must verify that the data you are planning to reuse was ethically collected or generated, was collected for a purpose in alignment with additional applications, and ensure that the study does not use the data irresponsibly or immorally. It is our duty as scientists and citizens to move science forward while protecting sensitive data and presenting studies fairly. Secondary analysis of public data requires an Institutional Review Board (IRB) exemption confirming it does not fall within the regulatory definition of human subject research (see federal regulations on human subjects research protections (45 CFR 46.101(2)(b))) and a full review otherwise (e.g., software as a medical device where you are using AI or machine learning for diagnosis). Special considerations need to be made for data reuse:

- Ethics by data type: The ethics of sharing and using data depend heavily on the data type. For example, cell lines, animal models, and microbial data are typically very low risk for privacy or ethical issues; that is not to say, however, that they have no risk (e.g., sharing pathogenic sequence data or alluding to locations for endangered species that could increase the risk of poaching) [24]. Ethical concerns are most common when working with human data.
- HIPAA, PHI, and patient de-identification: The Health Insurance Portability and Accountability Act (HIPAA) of 1996 Privacy Rule is a federal law regarding protected health information (PHI) for individuals and their access to their own health information, as well as the specific permissible use and disclosure of PHI with other organizations [25]. Any purposeful or accidental disclosure of PHI as a HIPAA violation can lead to hefty fines. Health data used in research needs to be de-identified in a manner that makes re-identification highly unlikely unless that is the willful goal of the project. Genomics data is particularly susceptible to re-identification due to the uniquely identifying nature of the data itself; particularly when coupled with geographic collection metadata such as collection site. Well-constructed data usage agreements and controlled-access repositories can mitigate re-identification risks (see Rule 3).
- Consent: Consent is an agreement between healthcare groups and participants that adheres to The Privacy Rule and HIPAA authorizations. For example, the GTEx Live Donor Informed Consent Template (BBRB-PM-0018 [26]) asserts exclusion of access to participant PHI that the generated data will be saved for many years, that it will be available to scientists around the world, and that data may be used broadly for medical research. It is important to ensure that data reuse does not violate any initially obtained consent.
- Ethics specific to how/where data were obtained: Data may be obtained directly from an individual lab, or queried from a private or public repository. Publicly available data that can be downloaded by anyone tends to be at lowest risk for violating privacy and are the easiest to access (i.e., TCGA [4] and GTEx [5] gene expression data). However, when data is received from another investigator through direct sharing or from a controlled-access

repository, it becomes your obligation to secure that data and not further share with unauthorized individuals [24].

- Ethical design for data reuse: When reusing data, ensure fairness and equality with your representation of the data, including but not limited to ancestry or sex. For example, genetic study participation has been disproportionately overrepresented by European descendants, where one study found that as of 2018, individuals in GWAS catalogs consisted of 78% European, 10% Asian, 2% African, 1% Hispanic, and <1% for all other races [27]. In GTEx as of 2022, 84.6% donors were of white origin, 12.9% African American origin, 1.3% Asian origin, 1.1% unknown and no statistic for Hispanic/Latino origin [5]. Additionally, sex differences impact every area of health and have been largely disregarded in study design and have also been heavily unbalanced, especially in pharmaceutical trials where women were previously excluded entirely due to potential pregnancies during trials [28]. Since then, women are now included in study designs, though sex specificity has not been accounted for leading to vastly more adverse drug reactions in women due to inappropriate drug and dose recommendations [29]. Careful consideration of these factors will lead to more rigorous and accurate results and avoid perpetuating these issues in research.

Rule 5: Plan for needed data storage and compute requirements

Ask yourself, is my data "genomical" in size [30]? It is if you're working in genomics, but regardless of your data, the amount of data being integrated in research is growing dramatically and the cost for storage and computation is at a rate not seen previously. Knowledge is power, and in this case, it is knowledge of the resources that will be needed *before* you need it that will benefit your work most. Data storage and computing hardware requirements should be determined and documented prior to downloading any data sets. This avoids potentially time consuming and expensive surprises by: (1) identifying gaps in the current infrastructure; and (2) allowing expert support staff an opportunity to investigate the viable solutions, where possible, in a timely manner. While needs vary by the individual situation and domain, here, we discuss various criteria to determine your needs. Data size, type, level of access, and security are the major considerations for where data needs to be stored. All invested parties who will retrieve, process, and consume the data need to be identified and involved in answering the following questions:

Data size:

- What is the estimated size of data to be retrieved?
- How many individual files are included?
- How much data are expected to be produced during processing?
- Do you need backups of the original or processed data?

Data type:

- Are the data types large or small?
- Are the file types binary or textual?
- Are the files compressed?

Access requirements:

- Who needs (or does not need) access to the data?
- What is the level of access required?

- How often will access be needed?
- How often will reanalysis be necessary?
- Do any of the users require (or prefer) a data sharing platform with a graphic interface for ease of access?
- Are there any institutional policies or approvals to be considered prior to granting access?
- Do external collaborators need to be provided access to the institutional setup?

Data security [31]:

- Does the data include PHI?
- Does the access need to be restricted?
- What is the minimum level of data security required?
- How will they be secured?
- Will the data need to be deleted after a certain time period or event? Why, when, and how?
- Are there any institutional policies or approvals to be considered?
- Who will supervise data security?
- How often (e.g., half-yearly, annually) will the adopted policies and implementations need to be reviewed and verified?

Answers to the above questions will facilitate discussions regarding optimal data storage options(s) and dictate if other institutional entities such as IT or the office of sponsored research need to be involved. Further, costs associated with data storage and reanalysis needs should also be considered when selecting storage locations. Commonly used data storage locations are personal computer(s), High Performance Computing (HPC) environments, and commercial cloud storage services (e.g., Dropbox, Box, AWS). Depending on the data, it may need to be split based on type and stored in a data-type-specific location best suited for its purpose. For example, large data such as FASTQ and VCF files may need to be stored at a location where they are accessible from an HPC environment for downstream processing, whereas spreadsheets and text documents may need to be stored in a cloud storage service (e.g., Box) to facilitate accessibility for non-computational team members.

Computing or hardware requirements will depend on the type of data and the processing planned [32]. A personal computer might suffice to process small data sets, but access to HPC or cloud computing (e.g., AWS, Microsoft Azure, Google Cloud) is often needed to process large data sets. HPCs and cloud computing offer several advantages such as fast processors, multicore chips, higher memory resources, and graphics processing units (GPU), which enable parallelization and scalability of a large number of jobs.

Rule 6: Know what you are downloading

As outlined in Rule 2, downstream analysis can be affected by how the data was collected and processed. This becomes an even greater problem when trying to integrate multiple public data sets, each collected and processed separately. So, be mindful of the processing differences between them. In recent years, a number of projects have been established that use and share common data processing pipelines across data collections, allowing researchers to more easily combine data for more complex analyses. For example, recount3 [33] is a uniformly processed resource of hundreds of thousands of human and mouse RNA-Seq data sets designed to

facilitate meta-analysis and cross-study comparisons. Another great resource is the Bgee [6] database that contains "normal," healthy, wild-type expression data across 52 different species thereby providing a comparable reference of gene expression by anatomical entities within and between species. A third example is BioDataome [34]. BioDataome has approximately 5,600 human and mouse expression and methylation data sets pre-processed in an analogous manner to allow for direct comparisons between data. All of these data repositories provide R packages facilitating programmatic data download (see Rule 7). Collections of equivalently processed data are becoming more of the norm, but data sets of interest may not be included in them. When this is the case, it may be necessary to reprocess the data to minimize the downstream impact of differences in upstream processing.

Rule 7: Download data programmatically; verify data integrity

Data downloads should be performed in a programmatic manner for consistency, scalability, and reliability. Several tools have been developed to provide direct programmatic access to large public databases (e.g., SRA [20], GEO [2], ENA [35], TCGA [4]) and they should be implemented when possible. One example of these tools is the SRA toolkit [36] that contains a number of commands linked to data downloads (e.g., "fasterq-dump" may be implemented to download FASTQ files and "vdb-validate" to check the integrity of SRA data sets). Additional examples include database-specific computational packages/libraries such as Bioconductor packages "recount3" [33] and "GenomicDataCommons" [37]. A key step in the data download process is validating the checksum (typically the MD5 hash) provided in the hosting database to verify data integrity prior to data analysis and interpretation. While these checks can be performed manually, some tools such as the Genomic Data Commons (GDC) Data Transfer Tool Client [37], will automatically validate MD5 checksums as part of the download process.

Depending on the number of files and tool functionality and design, customized scripts or workflows (e.g., Snakemake [38], NextFlow [39]) may be needed for scalability. When possible, publicly available workflows that adopt workflow management systems are recommended. One example is "fetchngs" [40] provided by nf-core [41]. This Nextflow pipeline [39] uses tools such as SRA-tools [36] to download FASTQ files and metadata based on a user-provided accession ID list. When performing downloads in this way, documenting the sample identifier and database sources becomes a critical step for reproducibility. For studies deposited in databases such as GEO [2], recording the GEO accession number along with sample identifiers may be sufficient; however, other data sources are updated on a continuous basis (e.g., Genotype-Tissue Expression (GTEx) Portal [5]). In such cases, the database version (e.g., GTEx Analysis Release V8—dbGaP Accession phs000424.v8.p2) and the date the download is performed should be recorded and reported in publications.

These practices should also be applied to data types beyond sample data including, for example, genome references (e.g., genome/transcriptome FASTA files, annotations files). Checksums from genomic reference files should be validated and details linked to the files should be recorded, including the FASTA file type (e.g., primary assembly or entire genomic sequence that includes assembly patches and haplotypes), database name, assembly name, and version or release number. The same principles should be applied to other reference files such as GTF/GFF [42].

Rule 8: Do properly cite data

As with any journal article/publication information and resources used, any data generated by other researchers should be credited and properly cited. This practice benefits the original data generator by providing a tangible demonstration of value and impact beyond the initial data

publication. Failing to cite the data source withholds credit from researchers in the same way that failing to cite a journal article does. Public data citations support better quality and more transparent science, making a compelling argument for other researchers to contribute their own data to public data repositories. This process also supports improved reproducibility and credibility for your own research.

When it comes to citing data, the field still lacks a gold standard. However, existing best practices include citing the original paper where the data was published or a data object identifier (DOI) generated from a persistent database like Zenodo [43] or figshare [44]. In addition, consortium projects such as GTEx [5] and TCGA [4] often provide guidance on citation practices for their repositories. In an effort to increase the citability of public data and establish the data itself as a scientific output of value separate from the associated manuscript, many researchers now publish data sets independently from their associated publications. Journals for this purpose, such as *Scientific Data*, now exist [45]. If there is no guideline associated with the data set in general the citation should include the generators, where the data was obtained from, accession numbers, the version number for the data (if applicable), and the date it was accessed.

Non-profit organizations such as DataCite and Crossref provide unique persistent identifiers (PIDs) to data sets to improve tracking usage and facilitate linking to the publication of origin [46]. Check and see if the public data you are using in your analysis has a PID that can be cited or included in the methods section. In the future, PIDs may perhaps be linked to an individual's ORCID number to provide a standardized data citation approach. A study looking at the correlation between if data was publicly available and the citation rate of the original paper demonstrated that those including publicly available data within the paper was associated with a 69% increase in citations [47]. Furthermore, investigators who share public data sets well have an increase in the impact of their own research. Articles with links to data repositories or that include PIDs are more highly cited than those without [24]. In summary, public data use benefits both the creator and the user.

Rule 9: Make your reprocessed data and models FAIR and share

All of the rules mentioned above are possible to adhere to because researchers made their data Findable, Accessible, Interoperable, and Reusable (FAIR) [48]. As a contributing member of the research ecosystem, you should pass it on, too! Make sure any additional data you generate, reuse, and integrate (i.e., data models), adheres to the FAIR principles. recount3 [33] and Bgee [6] are good examples of public data that have been reprocessed to correct for batch effects and made available under FAIR data principles. If your research is funded through the NIH, then you must adhere to a Data Management and Sharing Plan as outlined in the NIH policy starting in 2023 [10]. Recent research shows that only 6.8% of authors respond to requests for data sharing dramatically reducing the impact of most data and the knowledge to be gained from its use [49]. To be FAIR:

- Improve data findability: As highlighted in Rule 8, submit your data to stable open-access public data repositories that provide DOIs such as Zenodo [43] and figshare [44]. Personal, lab, group, or institute sites are not good long-term solutions. Socialize your data by sharing on social media platforms such as Twitter. Blogs and news articles on lab and/or institute sites and presenting at conferences where data are clearly identified with DOIs help others to discover useful data. By diversifying where and how you share with the community about your data, you cast the widest net in order to catch the attention of potential researchers who would also benefit from access to your data.

- Improve data accessibility: Open access to publications is important to science and plays a critical role in reproducing and advancing science. Making your data readily accessible is equally important. Share your raw as well as processed data so that the analysis performed can be reproduced fully and with minimal effort.
- Improve data interoperability: Interoperability refers to the ability of data from different sources to be able to integrate with minimal effort [48]. A good example is the Fast Health Interoperability Resources (FHIR) standard for health care data exchange [50]. This becomes even more critical in studies where data from different sources are being reused, so make sure the data you provide is highly interoperable by including relevant metadata and adhering to appropriate and reasonable file and data conventions within the field.
- Improve data reuse: Make your data reuse, reusable. Providing data in standard and popular formats goes a long way in making it reusable. Incomplete metadata and methods can severely limit the reuse (and usefulness) of data. Don't skimp by providing the bare minimum data and metadata needed to satisfy the requirements of the granting body, governing body, journal, etc., that you're looking to communicate your work through. Abstaining from providing all necessary assets to reproduce the work does a disservice to you, your colleagues, your lab and institute, and the scientific community. It is not just about the input and the output: There's a whole bunch of research, development, refinement, knowledge, and other work done in between data input and output that needs to be captured, codified, and shared with the work itself.

Rule 10: Make your processing, pipelines, and code FAIR

Tools and methods used for the analysis and interpretation of public data sets should also adhere to the FAIR guidelines for coding and software development [51]. Many of the FAIR principles for data (outlined in Rule 9) are directly applicable to software, but others require modification for application to software [52]. For example, persistent identifiers should be generated and recorded for novel pipelines, software, and research tools. Findability of software should be linked to the traceability of the source code under version control (e.g., GitHub [53], GitLab [54], BitBucket [55]) and reporting of appropriate metadata such as software versions. Similarly, software and code associated with an analysis should be accessible through repositories and (when appropriate) software archives such as CRAN [56], Bioconductor [57], PyPI [58], and Conda [59]. The same is true for software dependencies. Software containers (e.g., Docker [60], Singularity [61]) can also be implemented to enable software portability and analysis reproducibility. In the absence of containerized software, the necessary information for how to build and install a published tool should be provided.

Because the dynamic nature of software can mean the most up to date version is different from the version that was published, proper documentation and tagging (e.g., GitHub tags linked to released versions of software) allows researchers to find the exact package versions used during a study, therefore facilitating reproducible analyses. Continuous integration approaches (the software development practice of automating builds, static analysis, and tests on code changes that were pushed to a central repository) can automate time-consuming tasks associated with pipeline and software development. Similarly, static code analysis tools in continuous integration pipelines can help automate source code quality analysis allowing early identification of potential bugs, security vulnerabilities, performance issues, or deviations from the project/organization and coding guidelines.

Inclusion of continuous integration and static code analysis tools allow for rapid feedback loops in code inspection, reducing the time reviewers need to spend reviewing, and reducing

cost of time and funding of development maintenance [62,63]. Journals are increasingly requiring that the code used for analysis is made available [64]. In addition to ensuring that your processing, pipelines, and code are FAIR, the steps above will help to ensure code review is not an onerous task being performed during the submission process and your methods are reproducible for other scientists [65–67]. For more detailed information about how you can make your research more computationally reproducible, refer to [68].

Conclusion

In summary, biological data reuse is not only good for science, but also it is the right thing to do in order to extract the greatest societal impact from the samples and funding that patients, donors, and taxpayers generously provide. Here, we covered Ten Simple Rules for data reuse spanning the periods before, during, and after data download. This paper serves as a guide for both data users and generators in the community.

References

1. Navarro FCP, Mohsen H, Yan C, Li S, Gu M, Meyerson W, et al. Genomics and data science: an application within an umbrella. Genome Biol. 2019; 20:109. https://doi.org/10.1186/s13059-019-1724-1 PMID: 31142351
2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41:D991–5. https://doi.org/10.1093/nar/gks1193 PMID: 23193258
3. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39:1181–1186. https://doi.org/10.1038/ng1007-1181 PMID: 17898773
4. The Cancer Genome Atlas Program. In: National Cancer Institute [Internet]. 13 Jun 2018 [cited 2022 Jun 6]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
5. GTEx Portal. [cited 2022 Apr 26]. Available from: https://gtexportal.org/home/tissueSummaryPage.
6. Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, de Farias TM, et al. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Res. 2021; 49:D831–D847. https://doi.org/10.1093/nar/gkaa793 PMID: 33037820
7. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. elife. 2017:6. https://doi.org/10.7554/eLife.27041 PMID: 29206104
8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. https://doi.org/10.1038/nature11247 PMID: 22955616
9. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020; 48:D882–D889. https://doi.org/10.1093/nar/gkz1062 PMID: 31713622
10. Kozlov M. NIH issues a seismic mandate: share data publicly. Nature. 2022; 602:558–559. https://doi.org/10.1038/d41586-022-00402-1 PMID: 35173323
11. Boté J-J, Termens M. Reusing Data Technical and Ethical Challenges. DESIDOC Journal of Library & Information. Technology. 2019:329–337. https://doi.org/10.14429/djlit.39.06.14807
12. Parker M, Bull S. Sharing Public Health Research Data: Toward the Development of Ethical Data-Sharing Practice in Low- and Middle-Income Settings. J Empir Res Hum Res Ethics. 2015; 10:217–224. https://doi.org/10.1177/1556264615593494 PMID: 26297744
13. Organization WH. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. World Health Organization. 2022:p. v, 18 p.
14. Genomic Data Science Community Network. Diversifying the genomic data science research community. Genome Res. 2022. https://doi.org/10.1101/gr.276496.121 PMID: 35858750
15. Introduction. In: Alevin-fry requant [Internet]. [cited 2022 May 24]. Available from: https://combine-lab.github.io/quantaf/.
16. Boulesteix A-L, Hoffmann S, Charlton A, Seibold H. A replication crisis in methodological research? Significance. 2020; 17:18–21. https://doi.org/10.1111/1740-9713.01444

17. Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun. 2018; 9:2818. https://doi.org/10.1038/s41467-018-04976-1 PMID: 30026557
18. Kumuthini J, Chimenti M, Nahnsen S, Peltzer A, Meraba R, McFadyen R, et al. Ten simple rules for providing effective bioinformatics research support. PLoS Comput Biol. 2020; 16:e1007531. https://doi.org/10.1371/journal.pcbi.1007531 PMID: 32214318
19. Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M. Topic Group "Initial Data Analysis" of the STRATOS Initiative. Ten simple rules for initial data analysis. PLoS Comput Biol. 2022; 18:e1009819. https://doi.org/10.1371/journal.pcbi.1009819 PMID: 35202399
20. Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011; 39:D19–D21. https://doi.org/10.1093/nar/gkq1019 PMID: 21062823
21. Labastida I, Margoni T. Licensing FAIR data for reuse. Data Intellegence. 2020; 2:199–207. https://doi.org/10.1162/dint_a_00042
22. PLOS ONE. [cited 2022 Jun 13]. Available from: https://journals.plos.org/plosone/s/data-availability.
23. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. 2020; 52:646–654. https://doi.org/10.1038/s41588-020-0651-0 PMID: 32601475
24. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020; 21:615–629. https://doi.org/10.1038/s41576-020-0257-5 PMID: 32694666
25. Office for Civil Rights (OCR). Summary of the HIPAA Privacy Rule. [cited 2022 Apr 26]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.
26. GTEx Informed Consent Template. [cited 2022 Apr 26]. Available from: https://biospecimens.cancer.gov/resources/sops/library.asp.
27. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019; 177:26–31. https://doi.org/10.1016/j.cell.2019.02.048 PMID: 30901543
28. Oertelt-Prigione S, Mariman E. The impact of sex differences on genomic research. Int J Biochem Cell Biol. 2020; 124:105774. https://doi.org/10.1016/j.biocel.2020.105774 PMID: 32470538
29. Zucker I, Prendergast BJ. Sex differences in pharmacokinetics predict adverse drug reactions in women. Biol Sex Differ. 2020; 11:32. https://doi.org/10.1186/s13293-020-00308-5 PMID: 32503637
30. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015; 13:e1002195. https://doi.org/10.1371/journal.pbio.1002195 PMID: 26151137
31. Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol. 2016; 12:e1005097. https://doi.org/10.1371/journal.pcbi.1005097 PMID: 27764088
32. Brandies PA, Hogg CJ. Ten simple rules for getting started with command-line bioinformatics. PLoS Comput Biol. 2021; 17:e1008645. https://doi.org/10.1371/journal.pcbi.1008645 PMID: 33600404
33. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. bioRxiv. 2021. https://doi.org/10.1186/s13059-021-02533-6 PMID: 34844637
34. Lakiotaki K, Vorniotakis N, Tsagris M, Georgakopoulos G, Tsamardinos I. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. Database. 2018;2018. https://doi.org/10.1093/database/bay011 PMID: 29688366
35. EMBL-EBI. European Nucleotide Archive. [cited 2022 Jun 6]. Available from: https://www.ebi.ac.uk/ena/browser/home.
36. Sequence Read Archive Toolkit. [cited 2022 Jun 6]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
37. Morgan MT, Davis SR. GenomicDataCommons: a Bioconductor Interface to the NCI Genomic Data Commons. https://doi.org/10.1101/117200
38. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021; 10:33. https://doi.org/10.12688/f1000research.29032.2 PMID: 34035898
39. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017; 35:316–319. https://doi.org/10.1038/nbt.3820 PMID: 28398311
40. Patel H, Beber ME, Han DW, Ewels P, Espinosa-Carrasco J, Bot N-C, et al. nf-core/fetchngs: nf-core/fetchngs v1.5—Copper Cat. 2021. https://doi.org/10.5281/zenodo.5746702

41. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020; 38:276–278. https://doi.org/10.1038/s41587-020-0439-x PMID: 32055031
42. GFF3—GMOD. [cited 2022 Jun 6]. Available from: http://gmod.org/wiki/GFF3.
43. European Organization for Nuclear Research, OpenAIRE. Zenodo CERN. 2013. https://doi.org/10.25495/7GXK-RD71
44. Figshare. [cited 2022 Jun 6]. Available from: https://figshare.com/.
45. van den Berghe GJS-ASV, editor. Scientific Data. Nature Publishing Group. 2014-Current.
46. Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. In: Nature Publishing Group UK [Internet]. 4 Jun 2019 [cited 2022 Jun 6]. https://doi.org/10.1038/d41586-019-01715-4 PMID: 31164773
47. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS ONE. 2007; 2:e308. https://doi.org/10.1371/journal.pone.0000308 PMID: 17375194
48. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
49. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin Epidemiol. 2022. https://doi.org/10.1016/j.jclinepi.2022.05.019 PMID: 35654271
50. Index—FHIR v4.3.0. [cited 2022 Jun 10]. Available from: http://hl7.org/fhir/index.html.
51. Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Sci. 2020; 3:37–59. https://doi.org/10.3233/ds-190026
52. Jiménez RC, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. F1000Res. 2017:6. https://doi.org/10.12688/f1000research.11407.1 PMID: 28751965
53. Github. In: Github [Internet]. [cited 2022 Jun 6]. Available from: https://github.com/.
54. Gitlab. In: Gitlab [Internet]. [cited 2022 Jun 6]. Available from: https://about.gitlab.com/.
55. Bitbucket. In: Bitbucket [Internet]. [cited 2022 Jun 6]. Available from: https://bitbucket.org/product/.
56. The Comprehensive R Archive Network. [cited 2022 Jun 6]. Available from: https://cran.r-project.org/.
57. Bioconductor—Home. [cited 2022 Jun 6]. Available from: https://bioconductor.org/.
58. PyPI · The Python Package Index. In: PyPI [Internet]. [cited 2022 Jun 6]. Available from: https://pypi.org/.
59. Conda—Conda documentation. [cited 2022 Jun 6]. Available from: https://docs.conda.io/en/latest/.
60. Docker. [cited 2022 Jun 6]. Available from: https://www.docker.com/.
61. SingularityCE. In: Sylabs [Internet]. 31 Mar 2022 [cited 2022 Jun 6]. Available from: https://sylabs.io/singularity/.
62. Ferenc K, Otto K, de Oliveira Neto FG, López MD, Horkoff J, Schliep A. Empirical study on software and process quality in bioinformatics tools. bioRxiv. 2022:p. 2022.03.10.483804. https://doi.org/10.1101/2022.03.10.483804
63. Georgeson P, Syme A, Sloggett C, Chung J, Dashnow H, Milton M, et al. Bionitio: demonstrating and facilitating best practices for bioinformatics command-line software. Gigascience. 2019:8. https://doi.org/10.1093/gigascience/giz109 PMID: 31544213
64. Cadwallader L, Mac Gabhann F, Papin J, Pitzer VE. Advancing code sharing in the computational biology community. PLoS Comput Biol. 2022; 18:e1010193. https://doi.org/10.1371/journal.pcbi.1010193 PMID: 35653366
65. Hunter-Zinck H, de Siqueira AF, Vásquez VN, Barnes R, Martinez CC. Ten simple rules on writing clean and reliable open-source scientific software. PLoS Comput Biol. 2021; 17:e1009481. https://doi.org/10.1371/journal.pcbi.1009481 PMID: 34762641
66. Fungtammasan A, Lee A, Taroni J, Wheeler K, Chin C-S, Davis S, et al. Ten simple rules for large-scale data processing. PLoS Comput Biol. 2022; 18:e1009757. https://doi.org/10.1371/journal.pcbi.1009757 PMID: 35143491
67. Balaban G, Grytten I, Rand KD, Scheffer L, Sandve GK. Ten simple rules for quick and dirty scientific programming. PLoS Comput Biol. 2021; 17:e1008549. https://doi.org/10.1371/journal.pcbi.1008549 PMID: 33705383

68. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021; 18:1132–1135. https://doi.org/10.1038/s41592-021-01256-7 PMID: 34462593

D.3 Evolution

Further explanations for Chapter 3.

D.3.1 The origin of life

- Preiner et al. (2020) — Surveys the main competing scenarios (RNA world, metabolism first, compartments, vent and surface settings) and argues for combining them. A useful overview of the field as it stands after the classic debates. *(review; CC BY 4.0)*

  The Future of Origin of Life Research: Bridging Decades-Old Divisions. Preiner et al. (2020). *Life*. doi:10.3390/life10030020

  > Research on the origin of life is highly heterogeneous. After a peculiar historical development, it still includes strongly opposed views which potentially hinder progress. In the 1st Interdisciplinary Origin of Life Meeting, early-career researchers gathered to explore the commonalities between theories and approaches, critical divergence points, and expectations for the future. We find that even though classical approaches and theories-e.g. bottom-up and top-down, RNA world vs. metabolism-first-have been prevalent in origin of life research, they are ceasing to be mutually exclusive and they can and should feed integrating approaches. Here we focus on pressing questions and recent developments that bridge the classical disciplines and approaches, and highlight expectations for future endeavours in origin of life research.

- Pross and Pascal (2013) — A short, readable essay on which questions about the origin of life are answerable in principle and which are probably not, including the historical uncertainty of any scenario. *(perspective; CC BY (Europe PMC))*

  The origin of life: what we know, what we can know and what we will never know. Pross & Pascal (2013). *Open Biology*. doi:10.1098/rsob.120190

  > The origin of life (OOL) problem remains one of the more challenging scientific questions of all time. In this essay, we propose that following recent experimental and theoretical advances in systems chemistry, the underlying principle governing the emergence of life on the Earth can in its broadest sense be specified, and may be stated as follows: all stable (persistent) replicating systems will tend to evolve over time towards systems of greater stability. The stability kind referred to, however, is dynamic kinetic stability, and quite distinct from the traditional thermodynamic stability which conventionally dominates physical and chemical thinking. Significantly, that stability kind is generally found to be enhanced by increasing complexification, since added features in the replicating system that improve replication efficiency will be reproduced, thereby offering an explanation for the emergence of life's extraordinary complexity. On the basis of that simple principle, a fundamental reassessment of the underlying chemistry-biology relationship is possible, one with broad ramifications. In the context of the OOL question, this novel perspective can assist in clarifying central ahistoric aspects of abiogenesis, as opposed to the many historic aspects that have probably been forever lost in the mists of time.

D.3.2 Basic evolutionary concepts

- Aris-Brosou and Rodrigue (2019) — An introduction to the population-genetic and probabilistic foundations of molecular evolution, from mutation, selection and drift to substitution models. Bridges this section and the methods in the phylogeny chapter. *(book chapter; CC BY 4.0)*

  A Not-So-Long Introduction to Computational Molecular Evolution. Aris-Brosou & Rodrigue (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_3

> In this chapter, we give a not-so-long and self-contained introduction to computational molecular evolution. In particular, we present the emergence of the use of likelihood-based methods, review the standard DNA substitution models, and introduce how model choice operates. We also present recent developments in inferring absolute divergence times and rates on a phylogeny, before showing how state-of-the-art models take inspiration from diffusion theory to link population genetics, which traditionally focuses at a taxonomic level below that of the species, and molecular evolution. Although this is not a cookbook chapter, we try and point to popular programs and implementations along the way.

- Casillas and Barbadilla (2017) — A historical and conceptual review of molecular population genetics, covering the neutral and nearly neutral theories, linked selection and the interplay of drift and selection in genomic data. *(review; CC BY (Europe PMC))*

  Molecular Population Genetics. Casillas & Barbadilla (2017). *Genetics*. doi:10.1534/genetics.116.196493

  > Molecular population genetics aims to explain genetic variation and molecular evolution from population genetics principles. The field was born 50 years ago with the first measures of genetic variation in allozyme loci, continued with the nucleotide sequencing era, and is currently in the era of population genomics. During this period, molecular population genetics has been revolutionized by progress in data acquisition and theoretical developments. The conceptual elegance of the neutral theory of molecular evolution or the footprint carved by natural selection on the patterns of genetic variation are two examples of the vast number of inspiring findings of population genetics research. Since the inception of the field, Drosophila has been the prominent model species: molecular variation in populations was first described in Drosophila and most of the population genetics hypotheses were tested in Drosophila species. In this review, we describe the main concepts, methods, and landmarks of molecular population genetics, using the Drosophila model as a reference. We describe the different genetic data sets made available by advances in molecular technologies, and the theoretical developments fostered by these data. Finally, we review the results and new insights provided by the population genomics approach, and conclude by enumerating challenges and new lines of inquiry posed by increasingly large population scale sequence data.

D.3.3 Homologues, orthologues and gene duplication

- Birchler and Yang (2022) — Reviews what happens to genes after small-scale and whole-genome duplication, including the dosage-balance hypothesis. Examples are mostly from plants, but the concepts are general. *(review; CC BY 4.0)*

  The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Birchler & Yang (2022). *The Plant Cell*. doi:10.1093/plcell/koac076

  > Gene duplications have long been recognized as a contributor to the evolution of genes with new functions. Multiple copies of genes can result from tandem duplication, from transposition to new chromosomes, or from whole-genome duplication (polyploidy). The most common fate is that one member of the pair is deleted to return the gene to the singleton state. Other paths involve the reduced expression of both copies (hypofunctionalization) that are held in duplicate to maintain sufficient quantity of function. The two copies can split functions (subfunctionalization) or can diverge to generate a new function (neofunctionalization). Retention of duplicates resulting from doubling of the whole genome occurs for genes involved with multicomponent interactions such as transcription factors and signal transduction components. In contrast, these classes of genes are underrepresented in small segmental duplications. This complementary pattern suggests that the balance of interactors affects the fate of the duplicate pair. We discuss the different mechanisms that maintain duplicated genes, which may change over time and intersect.

- Glover et al. (2019) — A community review from the Quest for Orthologs consortium on orthology concepts, inference methods, benchmarking and applications; complements the Altenhoff et al. chapter already cited. *(review; CC BY 4.0)*

  Advances and Applications in the Quest for Orthologs. Glover et al. (2019). *Molecular Biology and Evolution*. doi:10.1093/molbev/msz150

> Gene families evolve by the processes of speciation (creating orthologs), gene duplication (paralogs), and horizontal gene transfer (xenologs), in addition to sequence divergence and gene loss. Orthologs in particular play an essential role in comparative genomics and phylogenomic analyses. With the continued sequencing of organisms across the tree of life, the data are available to reconstruct the unique evolutionary histories of tens of thousands of gene families. Accurate reconstruction of these histories, however, is a challenging computational problem, and the focus of the Quest for Orthologs Consortium. We review the recent advances and outstanding challenges in this field, as revealed at a symposium and meeting held at the University of Southern California in 2017. Key advances have been made both at the level of orthology algorithm development and with respect to coordination across the community of algorithm developers and orthology end-users. Applications spanned a broad range, including gene function prediction, phylostratigraphy, genome evolution, and phylogenomics. The meetings highlighted the increasing use of meta-analyses integrating results from multiple different algorithms, and discussed ongoing challenges in orthology inference as well as the next steps toward improvement and integration of orthology resources.

- Kellis et al. (2016a) — Lecture-note style text on gene trees versus species trees and reconciliation to infer duplications, losses, orthologues and paralogues. *(book chapter; CC BY-NC-SA (book level; page marked 'not declared') – link only)*

  Inferring Orthologs/Paralogs, Gene Duplication and Loss. Kellis et al. (2016). *Computational Biology: Genomes, Networks, Evolution (MIT 6.047, LibreTexts)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/27%3A_Phylogenomics_II/27.05%3A_Inferring_Orthologs_Paralogs_Gene_Duplication_and_Loss

D.3.4 Protein evolution

- Forslund et al. (2019) — Describes how domain architectures arise by fusion, fission, duplication and loss of domains, and how they are analysed computationally; matches the domain-evolution subsection closely. *(book chapter; CC BY 4.0)*

  Evolution of Protein Domain Architectures. Forslund et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_15

  > This chapter reviews current research on how protein domain architectures evolve. We begin by summarizing work on the phylogenetic distribution of proteins, as this will directly impact which domain architectures can be formed in different species. Studies relating domain family size to occurrence have shown that they generally follow power law distributions, both within genomes and larger evolutionary groups. These findings were subsequently extended to multi-domain architectures. Genome evolution models that have been suggested to explain the shape of these distributions are reviewed, as well as evidence for selective pressure to expand certain domain families more than others. Each domain has an intrinsic combinatorial propensity, and the effects of this have been studied using measures of domain versatility or promiscuity. Next, we study the principles of protein domain architecture evolution and how these have been inferred from distributions of extant domain arrangements. Following this, we review inferences of ancestral domain architecture and the conclusions concerning domain architecture evolution mechanisms that can be drawn from these. Finally, we examine whether all known cases of a given domain architecture can be assumed to have a single common origin (monophyly) or have evolved convergently (polyphyly). We end by a discussion of some available tools for computational analysis or exploitation of protein domain architectures and their evolution.

- Joy et al. (2016) — An educational topic page on ancestral reconstruction with parsimony, maximum likelihood and Bayesian methods, including the resurrection of ancestral proteins. *(tutorial; CC BY 4.0)*

  Ancestral Reconstruction. Joy et al. (2016). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1004763

- Oliveira and Deane (2017) — A short review of how correlated mutations in multiple sequence alignments are used to predict residue contacts, protein structures and interactions. *(review; CC BY 4.0)*

  Co-evolution techniques are reshaping the way we do structural bioinformatics. de Oliveira & Deane (2017). *F1000Research*. doi:10.12688/f1000research.11543.1

  > Co-evolution techniques were originally conceived to assist in protein structure prediction by inferring pairs of residues that share spatial proximity. However, the functional relationships that can be extrapolated from co-evolution have also proven to be useful in a wide array of structural bioinformatics applications. These techniques are a powerful way to extract structural and functional information in a sequence-rich world.

D.3.5 Detecting selection

- Kosiol and Anisimova (2019) — Overview of methods to detect selection from both divergence (codon models, dN/dS) and polymorphism data, with their assumptions. *(book chapter; CC BY 4.0)*

  Selection Acting on Genomes. Kosiol & Anisimova (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_12

  > Populations evolve as mutations arise in individual organisms and, through hereditary transmission, may become ”fixed” (shared by all individuals) in the population. Most mutations are lethal or have negative fitness consequences for the organism. Others have essentially no effect on organismal fitness and can become fixed through the neutral stochastic process known as random drift. However, mutations may also produce a selective advantage that boosts their chances of reaching fixation. Regions of genomes where new mutations are beneficial, rather than neutral or deleterious, tend to evolve more rapidly due to positive selection. Genes involved in immunity and defense are a well-known example; rapid evolution in these genes presumably occurs because new mutations help organisms to prevail in evolutionary ”arms races” with pathogens. In recent years genome-wide scans for selection have enlarged our understanding of the genome evolution of various species. In this chapter, we will focus on methods to detect selection on the genome. In particular, we will discuss probabilistic models and how they have changed with the advent of new genome-wide data now available.

- Jones et al. (2019) — Discusses how model misspecification and data problems can produce false signals of positive selection in codon-model analyses; relevant to the pitfalls subsection. *(book chapter; CC BY 4.0)*

  Looking for Darwin in Genomic Sequences: Validity and Success Depends on the Relationship Between Model and Data. Jones et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_13

  > Codon substitution models (CSMs) are commonly used to infer the history of natural section for a set of protein-coding sequences, often with the explicit goal of detecting the signature of positive Darwinian selection. However, the validity and success of CSMs used in conjunction with the maximum likelihood (ML) framework is sometimes challenged with claims that the approach might too often support false conclusions. In this chapter, we use a case study approach to identify four legitimate statistical difficulties associated with inference of evolutionary events using CSMs. These include: (1) model misspecification, (2) low information content, (3) the confounding of processes, and (4) phenomenological load, or PL. While past criticisms of CSMs can be connected to these issues, the historical critiques were often misdirected, or overstated, because they failed to recognize that the success of any model-based approach depends on the relationship between model and data. Here, we explore this relationship and provide a candid assessment of the limitations of CSMs to extract historical information from extant sequences. To aid in this assessment, we provide a brief overview of: (1) a more realistic way of thinking about the process of codon evolution framed in terms of population genetic parameters, and (2) a novel presentation of the ML statistical framework. We then divide the development of CSMs into two broad phases of scientific activity and show that the latter phase is characterized by increases in model complexity that can sometimes negatively impact inference of evolutionary mechanisms. Such problems are not yet widely appreciated by the users of CSMs. These problems can be avoided by using a model that is appropriate for the data; but, understanding the relationship between the data and a fitted model is a difficult task. We argue

> that the only way to properly understand that relationship is to perform in silico experiments using a generating process that can mimic the data as closely as possible. The mutation-selection modeling framework (MutSel) is presented as the basis of such a generating process. We contend that if complex CSMs continue to be developed for testing explicit mechanistic hypotheses, then additional analyses such as those described in here (e.g., penalized LRTs and estimation of PL) will need to be applied alongside the more traditional inferential methods.

- Kellis et al. (2016b) — Lecture-note text on genomic signatures of selection: dN/dS, Tajima's D, Fay and Wu's H, FST and haplotype-based sweep tests, with human examples. *(book chapter; CC BY-NC-SA (book level; page marked 'not declared') – link only)*

  Natural Selection. Kellis et al. (2016). *Computational Biology: Genomes, Networks, Evolution (MIT 6.047, LibreTexts)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/29%3A_Population_Genetic_Variation/29.04%3A_Natural_Selection

D.3.6 Genome evolution

- Koonin (2016) — Reviews the extent and mechanisms of horizontal gene transfer in prokaryotes and argues that it is essential for their long-term survival and evolvability. *(review; CC BY 4.0)*

  Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. Koonin (2016). *F1000Research*. doi:10.12688/f1000research.8737.1

  > The wide spread of gene exchange and loss in the prokaryotic world has prompted the concept of 'lateral genomics' to the point of an outright denial of the relevance of phylogenetic trees for evolution. However, the pronounced coherence congruence of the topologies of numerous gene trees, particularly those for (nearly) universal genes, translates into the notion of a statistical tree of life (STOL), which reflects a central trend of vertical evolution. The STOL can be employed as a framework for reconstruction of the evolutionary processes in the prokaryotic world. Quantitatively, however, horizontal gene transfer (HGT) dominates microbial evolution, with the rate of gene gain and loss being comparable to the rate of point mutations and much greater than the duplication rate. Theoretical models of evolution suggest that HGT is essential for the survival of microbial populations that otherwise deteriorate due to the Muller's ratchet effect. Apparently, at least some bacteria and archaea evolved dedicated vehicles for gene transfer that evolved from selfish elements such as plasmids and viruses. Recent phylogenomic analyses suggest that episodes of massive HGT were pivotal for the emergence of major groups of organisms such as multiple archaeal phyla as well as eukaryotes. Similar analyses appear to indicate that, in addition to donating hundreds of genes to the emerging eukaryotic lineage, mitochondrial endosymbiosis severely curtailed HGT. These results shed new light on the routes of evolutionary transitions, but caution is due given the inherent uncertainty of deep phylogenies.

- Puigbò et al. (2019) — Explains how comparing thousands of gene trees quantifies horizontal gene transfer and the remaining tree-like signal in prokaryotic evolution. *(book chapter; CC BY 4.0)*

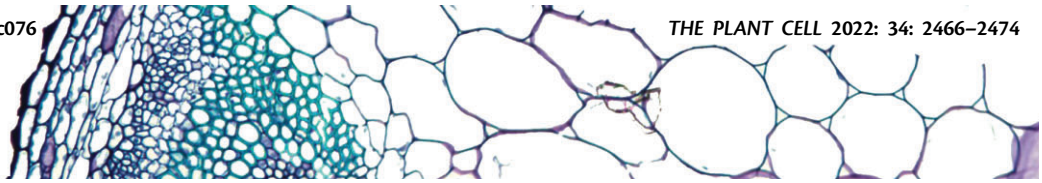
  Genome-Wide Comparative Analysis of Phylogenetic Trees: The Prokaryotic Forest of Life. Puigbò et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_8

  > Genome-wide comparison of phylogenetic trees is becoming an increasingly common approach in evolutionary genomics, and a variety of approaches for such comparison have been developed. In this article we present several methods for comparative analysis of large numbers of phylogenetic trees. To compare phylogenetic trees taking into account the bootstrap support for each internal branch, the boot-split distance (BSD) method is introduced as an extension of the previously developed split distance (SD) method for tree comparison. The BSD method implements the straightforward idea that comparison of phylogenetic trees can be made more robust by treating tree splits differentially depending on the bootstrap support. Approaches are also introduced for detecting treelike and netlike evolutionary trends in the phylogenetic Forest of Life (FOL), i.e., the entirety of the phylogenetic trees for conserved genes of prokaryotes. The principal method employed for this purpose includes mapping quartets of species onto trees to calculate the support of each quartet topology and so to quantify the tree and net contributions to the distances between species. We describe the applications methods used to analyze the FOL and the

results obtained with these methods. These results support the concept of the Tree of Life (TOL) as a central evolutionary trend in the FOL as opposed to the traditional view of the TOL as a ”species tree.”

The following open-access paper(s) are reproduced under their original licences.

https://doi.org/10.1093/plcell/koac076 THE PLANT CELL 2022: 34: 2466–2474



Perspective

The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation

James A. Birchler [1,†,*] and Hua Yang [1]

1 Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211, USA

*Author for correspondence: birchlerj@missouri.edu
†Senior author
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plcell) is: James A. Birchler (birchlerj@missouri.edu).

Abstract

Gene duplications have long been recognized as a contributor to the evolution of genes with new functions. Multiple copies of genes can result from tandem duplication, from transposition to new chromosomes, or from whole-genome duplication (polyploidy). The most common fate is that one member of the pair is deleted to return the gene to the singleton state. Other paths involve the reduced expression of both copies (hypofunctionalization) that are held in duplicate to maintain sufficient quantity of function. The two copies can split functions (subfunctionalization) or can diverge to generate a new function (neofunctionalization). Retention of duplicates resulting from doubling of the whole genome occurs for genes involved with multicomponent interactions such as transcription factors and signal transduction components. In contrast, these classes of genes are underrepresented in small segmental duplications. This complementary pattern suggests that the balance of interactors affects the fate of the duplicate pair. We discuss the different mechanisms that maintain duplicated genes, which may change over time and intersect.

Introduction

Mutations give rise to new variation and gene duplications give rise to new genes. New genes evolve from old genes. New genes foster the complexity of life leading to a multitude of new forms and functions. The trajectory of duplicate genes is usually a return to the singleton state, but retention can occur by various mechanisms. While there have been claims of completely de novo genes, they often include preexisting domains that are rearranged. Rearranged domains from different proteins, of course, can contribute to new functions. A truly de novo gene that is not selected against might be one that is antagonistic to a normal function that somehow confers an advantage to the organism. Some of the best candidates might be microRNAs due to their small size and modulatory action (Lu, 2019). Nevertheless, duplication is the predominant means of generating new gene functions. Here we discuss the various fates of duplicated genes.

Mechanisms of gene duplication (and deletion)

Gene duplications can occur in tandem via recombination between displaced repetitive sequences on homologues, via

Received December 06, 2021. Accepted February 17, 2022. Advance access publication March 7, 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of American Society of Plant Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

transposition by aberrant transposable element mobilizations (Zhang et al., 2013; Wang et al., 2015), or by being carried by Mutator-like elements (Jiang et al., 2004) or helitrons (Lal et al., 2003; Lai et al., 2005; Jameson et al., 2008; Barbbaglia et al., 2012), although the products are often gene fragments. Complete duplicates are in the minority but have the most consequences caused by the altered dosage (Zhang et al., 2021). Once a tandem duplication occurs, it can expand to a triplication or be reduced back to the singleton state by asymmetric recombination between the two members. Alternatively, duplicates can result from whole-genome duplication (WGD) (Wolfe and Shields, 1997; Simillion et al., 2002; Bowers et al., 2003; Blanc and Wolfe, 2004; Maere et al., 2005; Freeling, 2009; Jiao et al., 2011). In the plant kingdom, WGDs have occurred many times in various evolutionary lineages. Essentially, all angiosperms have a history of multiple WGDs of various ages (Jiao et al., 2011). Thus, whether a plant species is to be considered a polyploid or diploid depends on when one starts counting, blurring this distinction on the molecular level. The fate of most gene duplications is that one member of the pair is deleted or loses function, but some forces work to maintain the duplicate state.

In order for genes to be maintained in a genome, they must provide a function to the organism. Numerous types of mutational processes such as deletion via illegitimate recombination, ultraviolet light, replication errors, or transposable element insertion occur continuously, resulting in deletion or deterioration of genes unless purifying selection acts against the loss of the encoded function. For example, fish restricted to caves have become eyeless, losing the function of this otherwise critical organ when it is of little use (Yamamoto, 2004). Similarly, plants that have gained the ability to obtain a carbon source from other species often lose their chlorophyll (Wolfe et al., 1992). Thus, perhaps the most common fate of duplicated genes is that one member is deleted to return the gene copy status to singleton (Woodhouse et al., 2010).

Gene function is maintained through natural selection. If we consider the notion that duplicate genes are maintained so that a "back up" copy is present in the genome or that duplicates "buffer" each other, this would require that one copy of the duplicate pair would have to regularly fail to perform its function without being deleted, so that the "back up" copy could be selected upon. Early after a duplication event, a nullifying mutation in one copy of nondosage-sensitive genes would likely be of no effect and could persist with the single functional copy still present. Mutational assault shortly after WGD could be withstood better by the newly minted polyploid than the diploid progenitors (Innan and Kondrashov, 2010) because multiple copies of a gene are present, but this fact is unlikely to be a force for maintaining the duplicate state and will no longer occur as the polyploid genome fractionates back to a near diploid.

Another possible circumstance is that homoeologous copies of genes in an allopolyploid can become identical. One possible mechanism for this occurrence might be gene conversion, for which an example has been documented in yeast (Evangelisti and Conant, 2010). However, this is difficult to distinguish from homoeologous shuffling with nullisomic/tetrasomic compensation that will homogenize gene copies and can be clearly documented on the chromosomal level (Xiong et al., 2011). Homoeologous shuffling would homogenize segments of chromosomes in a syntenic relationship, which would allow a distinction.

To some degree, the fate of duplicate genes depends on the nature of the mutations that occur in the two copies that provide grist for selection. Muller (1932) coined terms for various types of mutations that have influenced the terminology in considering gene duplication changes from the progenitor (Table 1). The categories with regard to the fate of duplications are not always clear and indeed might change over the life of the duplication. Thus, there is a need for definition of terms, but the ultimate interest is the forces at play that affect the flux of the gene repertoire over time and how this influences the evolution of the genome.

Hypofunctionalization

A common trajectory of duplicated genes is the process of hypofunctionalization or dosage sharing (Qian et al., 2010; Gout and Lynch, 2015; Lan and Pritchard, 2016; Panchy et al., 2016; Veitia, 2017). In this case both members of a duplicate pair experience changes in expression level that reduce the total output (Figure 1). If the total amount of gene product is lowered to a level that is barely sufficient to supply the encoded function, then the duplicate pair would be maintained because the loss of either one of the two members of the pair would drop the total amount of gene expression below the threshold needed. The level of expression of the two members does not necessarily have to decline in parallel. Given that a certain level of gene expression for vital function differs across tissues, the members of the pair could diverge considerably in tissue-specific expression with only the critical tissues serving to maintain the level from both members to hold them in duplicate. Thus, the different expression levels of the duplicate pair are not necessarily a reflection of the partitioning of function when the hypofunctionalization is specific to a critical tissue. Another means by which hypofunctionalization might occur is if one member of a duplicate pair in a

Table 1 Types of mutations in gene duplicates (defined by Muller, 1932)

Mutation type	Phenotype
Amorphic	Complete loss of function
Hypomorphic	Partial loss of function
Hypermorphic	Over-expression
Antimorphic	Antagonistic relative to wild-type
Neomorphic	New function

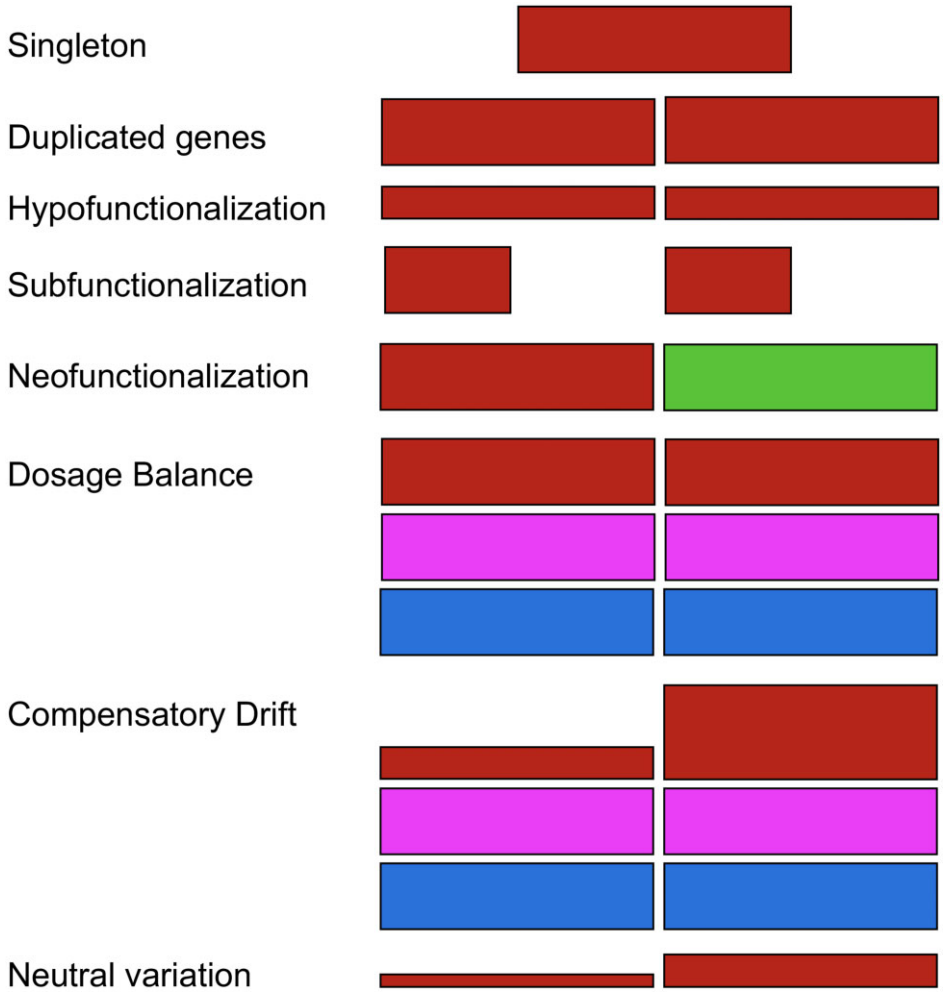


Figure 1 The fates of gene duplications. Starting from a singleton state a gene can become duplicated being then present in two copies via tandem duplication, transposed duplication, or WGD. Hypofunctionalization involves the reduction in expression of both copies of a duplicate pair to a threshold level at which both copies are needed for the specific function of the gene and thus both copies are maintained. Subfunctionalization is the division of function such that the two members of a pair confer only part of the functions of the progenitor singleton. Neofunctionalization is the diversification of one member of a duplicate pair for a new function. Dosage balance can mediate retention of specific classes of genes that are dosage sensitive to maintain their stoichiometric relationship with partners in multicomponent interactions. Nevertheless, compensatory drift refers to the observation that duplicates can diverge but depending on the total expression, a dosage balance effect will still operate with their partners. Neutral variation refers to the possibility that different members of a duplicate pair might have different expression levels under some circumstances in which the dosage or function is not an object of selection, similarly to apparent neutral variation of singleton genes.

path to deterioration becomes antagonistic to the remaining copy (antimorph) and causes a partially reduced total amount of function (Blackman et al., 2010; Panchy et al., 2016) near to the required threshold level but not below. This circumstance, however, would not necessarily result in the maintenance of both members of the duplication if the antimorphic mutation is deleted. Hypofunctionalization could be reversed if additional mutations increase the expression level of one member of the pair for sufficient function at the critical life cycle stage followed by deletion of the other copy. Hypofunctionalization could be a path to subfunctionalization or neofunctionalization, described below, depending on the outcome of additional mutations.

Neofunctionalization

The classical case of evolution by gene duplication involves circumstances in which duplicate genes diverge in their functions to provide new ones, an idea championed by Ohno (1970) (Figure 1). This process is referred to as neofunctionalization. When a new function arises that provides a successful evolutionary path, the duplicate pair of genes become fixed in the lineage. In this case, the pair of genes becomes maintained in duplicate essentially in perpetuity because the loss of either copy would be detrimental to the organism (Lynch and Conery, 2000).

The Drosophila bithorax complex is a classical example of neofunctionalization of duplicated genes, in which a set of homeobox genes is arranged in tandem to specify the

identity of different body parts to the segments in which they are expressed (Lewis, 1951). Lewis (1951) postulated that these genes arose by tandem duplication and then diverged in function so that each now specifies the different identities of the segmented body plan. This critical innovation arose early in animal development and is the foundation for the body plan of segmented animals. Humans, for example, have four different clusters of these tandemly arrayed genes, which are thought to have arisen by two WGD events in the evolutionary lineage. An early described example in plants involves the diverged tissue-specific functions of a duplicated anthocyanin gene in maize (Stadler, 1946; Stadler and Emmerling, 1956).

Genes affecting floral morphology and secondary metabolism provide excellent examples of neofunctionalization in plants (Chaanderbali et al., 2016). Different related transcription factors control various aspects of flower development and timing in evolutionary lineages (Sharma and Kramer, 2012; Madrid et al., 2021). The genes encoding cytochrome P450 genes are particularly prolific in plant species with diverse functions involved with hormones, lignin synthesis, secondary metabolites, defense compounds, and other biochemical functions. There are 256 members of this gene family in the Arabidopsis genome (Cannon et al., 2004). Some of these are organized as tandem arrays with individual members of the array presumably catalyzing different reactions. Genes involved with defense mechanisms are particularly prone to divergence of duplicate copies.

Freeling et al. (2015) posed the important question of what qualifies as a "new" function for a member of a duplicate pair of genes. Indeed, there is a tendency in the literature to assign neofunctionalization to differences from hypermorphic, antimorphic, or neomorphic mutations as defined by Muller (1932), but it is in fact difficult to determine if these cases are the basis of "new" genes or neutral variation at the evolutionary timeframe of examination. Muller suggested the definition of a neomorphic mutation as one for which an additional normal copy would have no impact on the phenotype. While this might be a good guide for defining neofunctionalization, genes that are dosage sensitive such as transcription factors and signaling components complicate even this definition.

Subfunctionalization

Subfunctionalization is related to neofunctionalization, but in this case, the duplicate pair partitions the functions of the progenitor gene (Figure 1). This process plays off the concept that genes typically will deteriorate and if there are two copies present, they may do so in different ways (Lynch and Conery, 2000). Thus, in order for all of the progenitor functions to be present, the duplicate pair must be retained in duplicate.

A classic example of subfunctionalization involves the hemoglobin cluster in humans (Hardison 2012). This oxygen-carrying protein has been duplicated by various means in vertebrate evolution. Some of the duplicated versions have taken on a different developmental timing but perform a similar function at the different stages. The copies with diverged functions are both required for normal function in the organism.

There is a tendency in the literature to ascribe subfunctionalization to any case in which duplicate genes have somewhat different expression patterns. Whether or not there has actually been a partitioning of function between the duplicate pair or whether this occurrence simply reflects the neutral variation of little to no consequence or hypofunctionalization is extremely difficult to assess (Panchy et al., 2016). Genes present in only a single copy in the genome can have allelic variation for the quantity or tissue-specific expression, which is likely evolutionarily neutral for the most part. Thus, when variation occurs for the quantity or tissue specificity of different members of a duplicate pair, this should not necessarily be ascribed to a partitioning of function. For duplicate genes of recent origin, it is particularly difficult to determine whether differences in expression reflect true subfunctionalization. The observation that most duplicate genes eventually transition back to a singleton state suggests that differences are not usually subfunctionalization. Duplicate genes that have persisted for millions of years with a maintenance of partitioned tissue specificity or other divided functions are candidates for subfunctionalization.

Dosage balance

Another force that acts to maintain duplicate genes is dosage balance (Birchler et al., 2005; Birchler and Veitia 2012, 2021). When whole-genome sequences became available, it was apparent that certain classes of genes tend to be found in duplicate more often than others (Blanc and Wolfe, 2004; Freeling and Thomas, 2006; Li et al., 2016; Kuzmin et al., 2020). In general, these tend to be genes involved in multicomponent interactions that are sensitive to gene dosage. Typical examples include genes that encode transcription factors, signal transduction components, or chromatin proteins (Birchler et al., 2001). Subsequently, microRNA genes were also found to be generally retained (Zhang et al., 2009; Abrouk et al., 2012; Zhao et al., 2015; Peterson et al., 2021), which is consistent with their frequent function in modulating transcription factor expression.

Dosage-sensitive genes are deemed to be responsible for the effects of aneuploidy given that they mimic the impact of aneuploidy on the phenotype of eukaryotic organisms (Birchler et al., 2001; 2005; Birchler and Veitia, 2021). Aneuploidy, or the change in dosage of individual chromosomes or chromosomal segments, is detrimental with both increases and decreases of chromatin. Changing the entire set of chromosomes has less of an effect, which has been recognized for a century (Blakeslee, 1921). The impact of aneuploidy on global gene expression is typically an inverse or proportional modulation across the genome relative to the varied chromosomal region (Birchler, 1979; Birchler and Newton, 1981; Guo and Birchler, 1994; Sun et al., 2013a, 2013b; Hou et al., 2018; Birchler and Veitia, 2021). The same general relationship occurs for individual dosage-sensitive

genes that are varied to the same copy number (Birchler et al., 2001; Birchler and Veitia, 2021). In contrast, less genome-wide modulation occurs from ploidy changes (Guo et al., 1996; Hou et al., 2018; Shi et al., 2021; Yang et al., 2021).

The preferential retention of dosage-sensitive genes was therefore postulated to be the result of negative fitness consequences if a member of an interacting group was deleted against the constant copy number of the interactors (Birchler et al., 2005; Freeling and Thomas, 2006; Veitia et al., 2008; 2013) (Figure 1). In other words, deletion of one member of a multicomponent interacting complex would be similar to an aneuploidy effect with similar detrimental consequences.

In a study of several plant species, Defoort et al. (2019) found that genes retained from WGD had a constraint for their level of expression. Genes retained from WGD showed an overrepresentation of involvement in protein–protein interactions whereas genes duplicated segmentally showed an underrepresentation of such interactions. In a study in Lotus with a single WGD (Shi et al., 2020), genes retained were the typical transcription factors and signaling components. These duplicates tended to have similar co-expression and a greater sequence conservation than segmental duplications. These studies support the concept that retention of selected classes of genes following WGD involves those associated with multicomponent interactions and that selection operates on variation that could upset the stoichiometry or balance of the components. Paralogs retained from WGD in rice and sorghum show fewer single-nucleotide polymorphisms than occur with tandem duplicates (Guo et al., 2019), illustrating the constraint on the former.

Certain classes of genes tend to return to single copy status through multiple polyploidy events (De Smet et al., 2013). This observation has been interpreted to indicate that these classes of genes have a tendency toward detrimental dominant mutations that would remove them if they have negative effects with regard to stoichiometric interactions in the cell.

Small-scale duplications

Analyses of classes of genes found in populations as segmental duplications reveal that those involved with multicomponent interactions are underrepresented, revealing a complementary pattern to those retained following WGD (Hakes et al., 2007; Freeling et al., 2008; Wang et al., 2011; Rodgers-Melnick et al., 2012; McGrath et al., 2014; Tasdighian et al., 2017). In this case, the increase of dosage-sensitive genes upon duplication would again mimic an aneuploid effect, be detrimental, and thus be selected against. The larger the size of the segmental duplication, the more likely it is to be detrimental. The most common effect of hyperploidy is an inverse modulation of global gene expression causing a reduced expression of a subset of genes across the genome. Larger segments of the genome that are varied in dosage tend to show a greater number of global modulations (Shi et al., 2021; Yang et al., 2021). This observation provides a molecular explanation for the detrimental effect of segmental duplication of dosage-sensitive genes.

Of course, if a small segmental duplication provides the organism with an exceptionally strong selective advantage, both copies might be retained, for example, if they confer disease resistance. Also, in some cases, absolute dosage can provide an advantageous increase of flux through a metabolic pathway that can be selected (Bekaert et al., 2011). The ultimate fate of a small segmental duplication will depend on the types and strength of selection operating on the increased copy number.

The complementary pattern of retained genes following WGD and segmental duplication is best explained by gene balance involving the stoichiometry of the interacting components of multicomponent cellular interactions, which if upset cause negative fitness consequences (Veitia 2002, 2003, 2004; Papp et al., 2003; Schuster-Bockler et al., 2010; Makanae et al., 2013; Robinson et al., 2021). Other potential explanations do not predict a complementary pattern of gene functions. Experimentally, the subtraction or addition of segments of the genome is more detrimental than a change in the whole genome (Hou et al., 2018; Shi et al., 2021; Yang et al., 2021), so the observed pattern of retained genes from WGD and segmental duplications provides a satisfying convergence between the evolutionary fate and experimental results.

Genome dominance

When WGDs are of allopolyploid origin, a typical result is that there is a difference in the overall expression level of the two genomes in the new configuration (Wang et al., 2006; Flagel et al., 2008; Garsmeur et al., 2014). This difference can persist long after formation of the polyploid (Murat et al., 2014; Woodhouse et al., 2014; Renny-Byfield et al., 2015; Cheng et al., 2016; Edger et al., 2017). Despite this difference, the retention via dosage balance appears to operate (Thomas et al., 2006; Woodhouse et al., 2010). However, as the retention via dosage balance deteriorates over time, members of the duplicate pair from the lesser expressed genome tend to be those that become deleted (Thomas et al., 2006; Schnable et al., 2011, 2012; Cheng et al., 2012), leaving the more highly expressed member as the remaining singleton. In a comparative study between maize and soybean, the degree of genome dominance was greater in maize than soybean. This difference was suggested to be due to the subgenomes being more distinct in the progenitor species to maize compared to soybean (Zhao et al., 2017). A greater number of homoeologs are retained in soybean than in maize. In soybean, genes in pericentromeric regions have higher levels of expression than their homoeologs in chromosome arms (Du et al., 2012).

Compensatory drift

While gene balance appears to be responsible for a large fraction of retained genes following WGD, the balance

constraints can be fluid. This conclusion is supported by the observation that longer timespans from WGDs to the present are associated with fewer remaining gene duplicate pairs (Maere et al., 2005; Aury et al., 2006). New effective stoichiometries or balance among interacting components must evolve to account for this observation (Birchler et al., 2005). Indeed, a recent study (Shi et al., 2020) shows that the degree of connectivity of retained genes corresponds to the degree of sequence divergence illustrating how stoichiometric constraints can evolve to new relationships. The more interactors a gene has, the less sequence divergence occurs between the duplicate members. Fewer interactors leads to less selection pressure on the pair of genes. The latter case has a greater potential to evolve new functional stoichiometries with its interactors because as the strength of the constraint decreases, the opportunity for new effective balances to emerge increases.

While dosage balance operates to maintain gene duplicates following WGD, this observation does not necessarily mean that the duplicate members retain similar expression levels. Modeling illustrates that the levels of expression of individual members of a duplicate pair can change without upsetting the overall balance (Thompson et al., 2016; Figure 1). An example of this situation is illustrated by the expression of the paralogs of the *FLC* flowering genes in polyploid Brassica (Calderwood et al., 2021). In this case, the individual genes have divergent expression levels, but the total amount is maintained. As noted above, variation in expression levels of individual members of a duplicate pair might be neutral in some tissues, while having a critical level in specific cell types. In plants, the haploid gametophyte generation is likely to be a situation in which the stoichiometric effects are more critical than in the diploid sporophyte, given that aneuploidy in haploids has stronger effects than in diploids (Shi et al., 2021; Yang et al., 2021). The competition to achieve fertilization would be a strong selective force. Indeed, homoeologs expressed in pollen appear to have longer retention times than those with other expression patterns, which has been attributed to the stronger selection on stoichiometric change in the haploid state (Chettor et al., 2014).

Neutral variation

Singleton genes can show significant variation in their levels of expression (Schadt et al., 2003; West et al., 2007; Zhou et al., 2020). This variation is likely to be essentially neutral, especially for nondosage-sensitive genes. While, as noted above, there are constraints on the level of expression of dosage-sensitive genes, there could also be neutral variation for the level of expression of these classes as well in certain cell types or tissues (Figure 1). Thus, variation in the level of expression of members of a duplicate pair could simply be neutral variation as occurs with singleton genes. The fact that the duplicates differ in expression is not necessarily a reflection of split or new functions. Thus, it is extremely difficult to classify the fate of duplicate genes on this basis alone.

Intersection of fates

The fate of duplicate genes and their impact to some degree depends on the time from the duplication event. Early after duplication a certain level of buffering is operative in that deletion or mutation of one member of a pair could have little impact on the organism compared to deletion or mutation of a singleton copy. Hypofunctionalization could be a major aspect somewhat later after the initial event as the two members deteriorate to some degree. Genome dominance in which one genome in an allopolyploid is expressed at a generalized higher level than the other can have an impact on the fractionation process with the lesser expressed genome losing more genes (Schnable et al., 2011; Edger et al., 2017). Dosage balance is in play immediately after duplication and can operate over millions of years but will likely decay with time depending on the degree of connectivity. Maintenance by dosage balance of transcription factors and signaling components for longer periods of evolutionary time than other classes of genes would provide a longer period of time for new functions and split functions to arise in the very classes of genes that would generate morphological innovation. It is difficult to determine the course of neo and subfunctionalization, which potentially could arise quickly after duplication, or after long periods during which the duplicate pair is maintained by other means, but those duplicates with very extended lives likely involve these two paths.

Funding

Research supported by National Science Foundation grant IOS-1545780.

Conflict of interest statement. None declared.

References

Abrouk M, Zhang R, Murat F, Li A, Point C, Mao L, Salse J (2012) Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole genome duplication. Plant Cell **24:** 1776–1792

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach O, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature **444:** 171–178

Barbbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC (2012) Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics **190**: 965–975

Bekaert M, Edger PP, Pires JC, Conant GC (2011) Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. Plant Cell **23:** 1151–1155

Birchler JA (1979) A study of enzyme activities in a dosage series of the long arm of chromosome one in maize. Genetics **92:** 1211–1229

Birchler JA, Bhadra U, Pal Bhadra M, Auger DL (2001) Dosage dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol **234:** 275–288

Birchler JA, Newton KJ (1981) Modulation of protein levels in chromosomal dosage series of maize: the biochemical basis of aneuploid syndromes. Genetics **99:** 247–266

Birchler JA, Riddle NC, Auger DL, Veitia RA (2005). Dosage balance in gene regulation: biological implications. Trends Genet **21:** 219–226

Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci USA **109:** 14746–14753

Birchler JA, Veitia RA (2021) One hundred years of gene balance: how stoichiometric issues affect gene expression, genome evolution, and quantitative traits. Cytogenet Genome Res **161**: 529–550 doi 10.1159/000519592

Blackman BK, Strasburg JL, Raduski AR, Michaels SD, Rieseberg LH (2010) The role of recently derived FT paralogs in sunflower domestication. Curr Biol **20:** 629–635

Blakeslee AF (1921) Types of mutations and their possible significance in evolution. Am Nat **55:** 254–267

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell **16:** 1679–1691

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

Calderwood A, Lloyd A, Hepworth J, Tudor EH, Jones DM, Woodhouse S, Bilham L, Chinoy C, Williams K, Corke F, et al. (2021) Total FLC transcript dynamics from divergent paralogue expression explains flowering diversity in *Brassica napus*. New Phytol **229:** 3534–3548

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol **4:** 10

Chaanderbali AS, Berger BA, Howarth DG, Soltis DE, Soltis PS (2016) Evolution of floral diversity: genomics, genes and gamma. Philos Trans B **372:** 20150509

Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, Cai C, Freeling M, Wang X (2016) Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*. New Phytologist **211:** 288–299

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X (2012) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PLoS One **7:** e36442

Chettor AM, Givan SA, Cole RA, Coker CT, Unger-Wallace E, Vejlupkova Z, Vollbrecht E, Fowler JE, Evans MM (2014) Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol **15:** 414

Defoort J, Van de Peer Y, Carretero-Paulet L (2019) The evolution of gene duplicates in angiosperms and the impact of protein-protein interactions and the mechanism of duplication. Genome Biol Evol **11:** 2292–2305

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA **110**: 2898–2903

Du J, Tian Z, Sui Y, Zhao M, song Q, Cannon SB, Cregan P, Ma J (2012) Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in paleopolyploid soybean. Plant Cell **24:** 21–32

Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al. (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell **29:** 2150–2167

Evangelisti AM, Conant GC (2010) Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. Genome Biol Evol **2:** 826–834

Flagel L, Udall J, Nettleton D, Wendel J (2008) Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. BMC Biol **6:** 16

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol **60**: 433–453

Freeling, M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D (2008) Many or most genes in *Arabidopsis* transposed after the origin of the order *Brassicales*. Genome Res **18:** 1924–1937

Freeling M, Scanlon MJ, Fowler JE (2015) Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. Curr Opin Genet Dev **35:** 110–118

Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16**: 805–814

Garsmeur O, Schable JC, Almeida A, Jourda C, D'Hont A, Freeling M (2014) Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol **31:** 448–454

Gout JF, Lynch M (2015) Maintenance and loss of duplicated genes by dosage subfunctionalization. Mol Biol Evol **32:** 2141–2148

Guo H, Jiao Y, Tan X, Wang X, Huang X, Jin H, Paterson AH (2019) Gene duplication and genetic innovation in cereal genomes. Genome Res **29:** 261–269

Guo M, Birchler JA (1994) *Trans*-acting dosage effects on the expression of model gene systems in maize aneuploids. Science **266:** 1999–2002

Guo M, Davis D, Birchler JA (1996) Dosage effects on gene expression in a maize ploidy series. Genetics **142:** 1349–1355

Hakes L, Pinney JW, Lovell Sc, Oliver SG, Robertson DL (2007) All duplicates are not equal: the difference between small-scale and genome duplications. Genome Biol **8:** R209

Hardison RC (2012) Evolution of hemoglobin and its genes. Cold Spring Harbor Perspect Med **2:** a011627

Hou J, Shi X, Chen C, Islam MS, Johnson AF, Kanno T, Huettel B, Yen MR, Hsu FM, Ji T, Chen, PY, et al. (2018) Global impacts of chromosomal imbalance on gene expression in Arabidopsis and other taxa. Proc Natl Acad Sci USA **115**: E11321–E11330

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet **11:** 97–108

Jameson N, Georgelis N, Fouladbash E, Martens S, Hannah LC (2008) Helitron mediated amplification of cytochrome P450 monooxygenase gene in maize. Plant Mol Biol **67:** 295–304

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature **431:** 569–573

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100

Kuzmin E, VanderSluis B, Nyugen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, et al. (2020) Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science **368:** eaaz5667

Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA **102:** 9068–9073

Lal S, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a helitron insertion. Plant Cell **15:** 381–391

Lan X, Pritchard JK (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science **352:** 1009–1013

Lewis EB (1951) Pseudoallelism and gene evolution. Cold Spring Harbor Symp Quant Biol **16:** 159–174

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R (2016) Gene duplicability of core genes is highly consistent across all Angiosperms. Plant Cell **28:** 326–344

Lu S (2019) De novo origination of MIRNAs through the generation of short inverted repeats in target genes. RNA Biol **16:** 846–859

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155

Madrid E, Severing E, de Ansorena E, Kiefer C, Brand L, Martinez-Gallegos R, Woetzel S, Kemi U, Jiao WB, Schneeberger K, et al. (2021) Transposition and duplication of MADS-domain transcription factor genes in annual and perennial Arabis species modulates flowering. Proc Natl Acad Sci USA **118:** e2109204118

Makanae K, Kiintaka R, Makino T, Kitano H, Moriya H (2013) Identification of dosage-sensitive genes in *Saccharomyces cerevisiae* using the genetic tug-of-war method. Genome Res **23:** 300–311

Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, de Peer YV (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102**: 5454–5459

McGrath CL, Gout JF, Johri P, Doak TG, Lynch M (2014) Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res **24:** 1665–1675

Muller HJ (1932) Further studies on the nature and causes of gene mutations. Proceedings of the Sixth International Congress of Genetics, pp 213–255

Murat F, Zhang R, Guizard S, Flores R, Armero A, Pont C, Steinbach D, Quesneville H, Cooke R, Salse J (2014) Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16 K protogenes. Genome Biol Evol **6**: 12–33

Ohno S (1970) Evolution by Gene Duplication. Springer, Berlin, Germany

Peterson KJ, Beavan A, Chabot PJ, McPeek MA, Pisani D, Fromm B, Simakov O (2021) microRNAs as indicators into the causes and consequences of whole genome duplication events. Mol Biol Evol **39**: msab344 doi 10.1093/molbev/msab344

Panchy N, Lehti-Shiu M, Shiu SH (2016) Evolution of gene duplication in plants. Plant Physiol **171:** 2294–2316

Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature **424:** 194–197

Qian W, Liao BY, Chang AYF, Zhang J (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet **26:** 425–430

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. Mol Biol Evol **32:** 1063–1071

Robinson D, Place M, Hose J, Jochem A, Gasch AP (2021) Natural variation in the consequences of gene overexpression and its implications for evolutionary trajectories. eLife **10:** e70564

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP (2012) Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome Res **22:** 95–105

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. (2003) Genetics of gene expression in surveyed maize, mouse and man. Nature **422:** 297–302

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA **108:** 4069–4074

Schnable JC, Wang X, Pires JC, Freeling M (2012) Escape from preferential retention following repeated whole genome duplications in plants. Front Plant Sci **3:** 94

Schuster-Bockler B, Conrad D, Bateman A (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One **5:** e9474

Sharma B, Kramer E (2012) Sub- and neo-functionalization of APETALA3 paralogs have contributed to the evolution of novel floral organ identity in Aguilegia (columbine, Ranunculaceae). New Phytologist **197:** 949–957

Shi T, Seyed Rahmani R, Gugger PF, Wang M, Li H, Zhang Y, Li Z, Wang Q, Van de Peer Y, Marchal K, et al. (2020) Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. Mol Biol Evol **37:** 2394–2413

Shi X, Yang H, Chen C, Hou J, Hanson KM, Albert PS, Ji T, Cheng J, Birchler JA (2021) Genomic imbalance determines positive and negative modulation of gene expression in diploid maize. Plant Cell **33:** 917–939

Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci USA **99**: 13627–13632

Stadler LJ (1946) Spontaneous mutation at the R locus in maize: I. The aleurone-color and plant color effects. Genetics **31**: 377–394

Stadler LJ, Emmerling MH (1956) Relation of unequal crossing over to the interdependence of the *Rr* elements (P) and (S). Genetics **41**: 124–137

Sun L, Johnson AF, Donohue RC, Li J, Cheng J, Birchler JA (2013a) Dosage compensation and inverse effects in triple X metafemales of *Drosophila*. Proc Natl Acad Sci USA **110:** 7383–7388

Sun L, Johnson AF, Li J, Lambdin AS, Cheng J, Birchler JA (2013b) Differential effect of aneuploidy on the X chromosome and genes with sex-biased expression in *Drosophila*. Proc Natl Acad Sci USA **110:** 16514–16519

Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S (2017) Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. Plant Cell **29:** 2766–2785

Thomas BC, Pedersen B, Freeling, M (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res **16**: 934–946

Thompson A, Zakon HH, Kirkpatrick M (2016) Compensatory drift and the evolutionary dynamics of dosage-sensitive duplicate genes. Genetics **202:** 765–774

Veitia RA (2002) Exploring the etiology of haploinsufficiency. Bioessays **24:** 175–184

Veitia RA (2003) Nonlinear effects in macromolecular assembly and dosage sensitivity. J Theor Biol **220:** 19–25

Veitia RA (2004) Gene dosage balance in cellular pathways: Implications for dominance and gene duplicability. Genetics **168:** 569–574

Veitia RA (2017) Gene duplicates: agents of robustness or fragility? Trends Genet **33:** 377–379

Veitia RA, Bottani S, Birchler JA (2008) Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet **24:** 390–397

Veitia RA, Bottani S, Birchler JA (2013) Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. Trends Genet **29**: 385–393

Wang J, Tian L, Lee HS, Wei NE, Jiang H, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, et al. (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172**: 507–517

Wang D, Yu C, Zuo T, Zhang J, Weber DF, Peterson T (2015) Alternative transposition generates new chimeric genes and segmental duplications at the maize *p1* locus. Genetics **201**: 925–935

Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Pateson AH (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS One **6:** e28150

West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St. Clair DA (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics **175:** 1441–1450

Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA **89:** 10648–10652

Wolfe KH, Shields, DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708–713

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc Natl Acad Sci USA **111:** 5283–5288

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol **8:** e1000409

Xiong Z, Gaeta RT, Pires JC (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. Proc Natl Acad Sci USA **108**: 7908–7913

Yamamoto Y (2004) Cavefish. Curr Biol **14:** R943

Yang H, Shi X, Chen C, Hou J, Ji T, Cheng J, Birchler JA (2021) Predominantly inverse modulation of gene expression in genomically unbalanced disomic haploid maize. Plant Cell **33:** 901–916

Zhang D, Leng L, Chen C, Huang J, Zhang Y, Yuan H, Ma C, Chen H, Zhang YE(2021) Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in Drosophila and humans. Nat Ecol Evol doi 10.1038/s41559-021-01614-w (December 30, 2021)

Zhang J, Zuo T, Pererson T (2013) Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. PLoS Genet **9:** e1003691

Zhang L, Chia JM, Kumari S, Stein JC, Liu Z, Marechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genet **5:** e1000716

Zhao M, Meyers BC, Cai C, Xu W, Ma J (2015) Evolutionary patterns and coevolutionary consequences of MIRNA genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean. Plant Cell **27:** 546–562

Zhao M, Zhang B, Lisch D, Ma J (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell **29:** 2974–2994

Zhou P, Li Z, Magnusson E, Cano FG, Crisp PA, Noshay JM, Grotewold E, Hirsch CN, Briggs SP, Springer NM (2020) Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. Plant Cell **32**: 1377–1396

PLOS | COMPUTATIONAL BIOLOGY

TOPIC PAGE

Ancestral Reconstruction

Jeffrey B. Joy[1,2], Richard H. Liang[1], Rosemary M. McCloskey[1], T. Nguyen[1], Art F. Y. Poon[1,2]*

1 BC Centre for Excellence in HIV/AIDS, Vancouver, British Columbia, Canada, **2** University of British Columbia, Department of Medicine, Vancouver, British Columbia, Canada

* apoon@cfenet.ubc.ca

CrossMark
click for updates

OPEN ACCESS

Citation: Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY (2016) Ancestral Reconstruction. PLoS Comput Biol 12(7): e1004763. doi:10.1371/journal.pcbi.1004763

Published: July 12, 2016

Copyright: © 2016 Joy et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors of this review were supported in part by an operating grant from the Canadian Institutes for Health Research awarded to AFYP (CIHR HOP111406). In addition, AFYP is supported by a Career Investigator Scholar Award from the Michael Smith Foundation for Health Research, St. Paul's Hospital Foundation and the Providence Health Care Research Institute, and by a CIHR New Investigator Award under the Canadian HIV Vaccine Initiative (CHVI) for Vaccine Discovery and Social Research. RMM was supported by a Vice President Research Undergraduate Student Research Award (VPR USRA) from Simon Fraser University. This work was partly supported by the Bill and Melinda Gates Foundation Award Number OPP1110049. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Ancestral reconstruction is the extrapolation back in time from measured characteristics of individuals (or populations) to their common ancestors. It is an important application of phylogenetics, the reconstruction and study of the evolutionary relationships among individuals, populations, or species to their ancestors. In the context of biology, ancestral reconstruction can be used to recover different kinds of ancestral character states, including the genetic sequence (ancestral sequence reconstruction), the amino acid sequence of a protein, the composition of a genome (e.g., gene order), a measurable characteristic of an organism (phenotype), and the geographic range of an ancestral population or species (ancestral range reconstruction). Nonbiological applications include the reconstruction of the vocabulary or phonemes of ancient languages [1] and cultural characteristics of ancient societies such as oral traditions [2] or marriage practices [3].

Ancestral reconstruction relies on a sufficiently realistic model of evolution to accurately recover ancestral states. No matter how well the model approximates the actual evolutionary history, however, one's ability to accurately reconstruct an ancestor deteriorates with increasing evolutionary time between that ancestor and its observed descendants. Additionally, more realistic models of evolution are inevitably more complex and difficult to calculate. Progress in the field of ancestral reconstruction has relied heavily on the exponential growth of computing power and the concomitant development of efficient computational algorithms (e.g., a dynamic programming algorithm for the joint maximum likelihood [ML] reconstruction of ancestral sequences [4]). Methods of ancestral reconstruction are often applied to a given phylogenetic tree that has already been inferred from the same data. While convenient, this approach has the disadvantage that its results are contingent on the accuracy of a single phylogenetic tree. In contrast, some researchers advocate a more computationally intensive Bayesian approach that accounts for uncertainty in tree reconstruction by evaluating ancestral reconstructions over many trees [5].

History

The concept of ancestral reconstruction is often credited to Emile Zuckerkandl and Linus Pauling. Motivated by the development of techniques for determining the primary (amino acid) sequence of proteins by Frederick Sanger in 1955 [6], Pauling and Zuckerkandl postulated [7] that such sequences could be used to infer not only the phylogeny relating the observed protein sequences but also the ancestral protein sequence at the earliest point (root) of this tree. However, the idea of reconstructing ancestors from measurable biological characteristics had already been developing in the field of cladistics, one of the precursors of modern

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: BEAST, Bayesian Evolutionary Analysis by Sampling Trees; BiSSE, binary state speciation and extinction model; COT, center-of-tree; DEC, dispersal-extinction-cladogenesis; DIVA, dispersal-vicariance analysis; FP, Fitch parsimony; HIV, human immunodeficiency virus; MCMC, Markov chain Monte Carlo; ML, maximum likelihood; SM, stochastic mapping; UPGMA, unweighted pair group method with arithmetic mean.

phylogenetics. Cladistic methods, which appeared as early as 1901, infer the evolutionary relationships of species on the basis of the distribution of shared characteristics, of which some are inferred to be descended from common ancestors. Furthermore, Theodoseus Dobzhansky and Alfred Sturtevant articulated the principles of ancestral reconstruction in a phylogenetic context in 1938, when inferring the evolutionary history of chromosomal inversions in *Drosophila pseudoobscura* [8]. Thus, ancestral reconstruction has its roots in several disciplines. Today, computational methods for ancestral reconstruction continue to be extended and applied in a diversity of settings so that ancestral states are being inferred not only for biological characteristics and the molecular sequences but also for the structure of folded proteins [9,10], the geographic location of populations and species (phylogeography) [11,12], and the higher-order structure of genomes [13].

Methods and Algorithms

Any attempt at ancestral reconstruction begins with a phylogeny. In general, a phylogeny is a tree-based hypothesis about the order in which populations (referred to as taxa) are related by descent from common ancestors. Observed taxa are represented by the tips or terminal nodes of the tree that are progressively connected by branches to their common ancestors, which are represented by the branching points of the tree that are usually referred to as the ancestral or internal nodes. Eventually, all lineages converge to the most recent common ancestor of the entire sample of taxa. In the context of ancestral reconstruction, a phylogeny is often treated as though it were a known quantity (with Bayesian approaches being an important exception). Because there can be an enormous number of phylogenies that are nearly equally effective at explaining the data, reducing the subset of phylogenies supported by the data to a single representative, or point estimate, can be a convenient and sometimes necessary simplifying assumption. Ancestral reconstruction can be thought of as the direct result of applying a hypothetical model of evolution to a given phylogeny. When the model contains one or more free parameters, the overall objective is to estimate these parameters on the basis of measured characteristics among the observed taxa (sequences) that descended from common ancestors. Parsimony is an important exception to this paradigm: though it has been shown that there are mathematical models for which it is the ML estimator [14], at its core, it is simply based on the heuristic that changes in character state are rare, without attempting to quantify that rarity.

Maximum Parsimony

Parsimony, known colloquially as "Occam's razor," refers to the principle of selecting the simplest of competing hypotheses. In the context of ancestral reconstruction, parsimony endeavors to find the distribution of ancestral states within a given tree that minimizes the total number of character state changes that would be necessary to explain the states observed at the tips of the tree. This method of maximum parsimony) [15] is one of the earliest formalized algorithms for reconstructing ancestral states. Maximum parsimony can be implemented by one of several algorithms. One of the earliest examples is Fitch's method [16], which assigns ancestral character states by parsimony via two traversals of a rooted binary tree. The first stage is a postorder traversal that proceeds from the tips toward the root of a tree by visiting descendant (child) nodes before their parents. Initially, we are determining the set of possible character states S_i for the i-th ancestor based on the observed character states of its descendants. Each assignment is the set intersection) of the character states of the ancestor's descendants; if the intersection is the empty set, then it is the set union). In the latter case, it is implied that a character state change has occurred between the ancestor and one of its two immediate descendants. Each such event counts towards the algorithm's cost function, which may be used to discriminate

among alternative trees on the basis of maximum parsimony. Next, a preorder traversal of the tree is performed, proceeding from the root towards the tips. Character states are then assigned to each descendant based on which character states it shares with its parent. Since the root has no parent node, one may be required to select a character state arbitrarily, specifically when more than one possible state has been reconstructed at the root. For example, consider a phylogeny recovered for a genus of plants containing six species, A–F (Fig 1), where each plant is pollinated by either a "bee," "hummingbird," or "wind." One obvious question is what the pollinators at deeper nodes were in the phylogeny of this genus of plants. Under maximum parsimony, an ancestral state reconstruction for this clade reveals that "hummingbird" is the most parsimonious ancestral state for the lower clade (plants D, E, F), that the ancestral states for the nodes in the top clade (plants A, B, C) are equivocal, and that both "hummingbird" or "bee" pollinators are equally plausible for the pollination state at the root of the phylogeny, supposing we have strong evidence from the fossil record that the root state is "hummingbird." Resolution of the root to "hummingbird" would yield the pattern of ancestral state reconstruction depicted by the symbols at the nodes (Fig 1) with the state requiring the fewest number of changes circled. Parsimony methods are intuitively appealing and highly efficient, such that they are still used in some cases to seed ML optimization algorithms with an initial phylogeny [17]. However, they suffer from several issues:

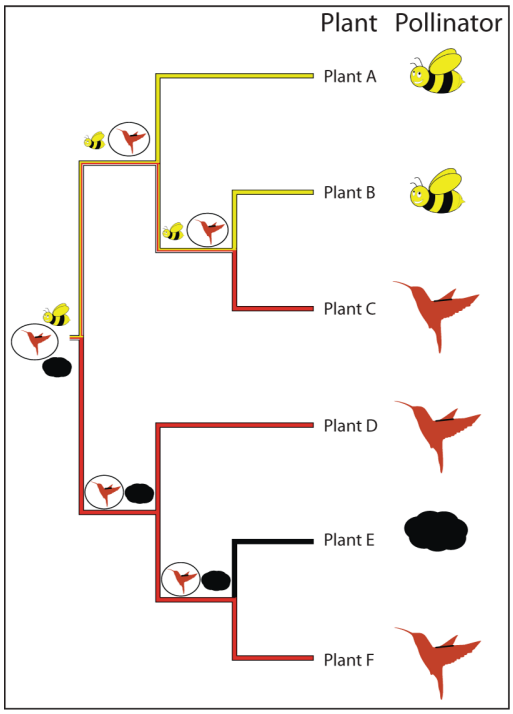


Fig 1. Phylogeny of a hypothetical genus of plants with pollination states of either "bees", "hummingbirds", or "wind" denoted by pictues at the tips. Pollination state nodes in the phylogenetic tree inferred under maximum parsimony are coloured on the branches leading into them (yellow represents "bee" pollination, red representing "hummingbird" pollination, and black representing "wind" pollination, dual coloured branches are equally parsimonious for the two states coloured). Assignment of "hummingbird" as the root state (because of prior knowledge from the fossil record) leads to the pattern of ancestral states represented by symbols at the nodes of the phylogeny, the state requiring the fewest number of changes to give rise to the pattern observed at the tips is circled at each node.

doi:10.1371/journal.pcbi.1004763.g001

1. Variation in rates of evolution. Fitch's method assumes that changes between all character states are equally likely to occur; thus, any change incurs the same cost for a given tree. This assumption is often unrealistic and can limit the accuracy of such methods. For example, transitions) tend to occur more often than transversions in the evolution of nucleic acids. This assumption can be relaxed by assigning differential costs to specific character state changes, resulting in a weighted parsimony algorithm [18].
2. Rapid evolution. The upshot of the "minimum evolution" heuristic underlying such methods is that such methods assume that changes are rare and thus are inappropriate in cases where change is the norm rather than the exception [19,20].
3. Variation in time among lineages. Parsimony methods implicitly assume that the same amount of evolutionary time has passed along every branch of the tree. Thus, they do not account for variation in branch lengths in the tree, which are often used to quantify the passage of evolutionary or chronological time. This limitation makes the technique liable to infer that one change occurred on a very short branch rather than multiple changes occurring on a very long branch, for example [21]. This shortcoming is addressed by model-based methods (both ML and Bayesian methods) that infer the stochastic process of evolution as it unfolds along each branch of a tree [22].
4. Statistical justification. Without a statistical model underlying the method, its estimates do not have well-defined uncertainties [19,21,23].

ML

ML methods of ancestral state reconstruction treat the character states at internal nodes of the tree as parameters and attempt to find the parameter values that maximize the probability of the data (the observed character states) given the hypothesis (a model of evolution and a phylogeny relating the observed sequences or taxa). Some of the earliest ML approaches to ancestral reconstruction were developed in the context of genetic sequence evolution [24,25]; similar models were also developed for the analogous case of discrete character evolution [26].

These approaches employ the same probabilistic framework as used to infer the phylogenetic tree [27]. In brief, the evolution of a genetic sequence is modelled by a time-reversible continuous time Markov process. In the simplest of these, all characters undergo independent state transitions (such as nucleotide substitutions) at a constant rate over time. This basic model is frequently extended to allow different rates on each branch of the tree. In reality, mutation rates may also vary over time (due, for example, to environmental changes); this can be modelled by allowing the rate parameters to evolve along the tree, at the expense of having an increased number of parameters. A model defines transition probabilities from states i to j along a branch of length t (in units of evolutionary time). The likelihood of a phylogeny is computed from a nested sum of transition probabilities that corresponds to the hierarchical structure of the proposed tree. At each node, the likelihood of its descendants is summed over all possible ancestral character states at that node:

$$L_x = \sum_{S_x \in \Omega} P(S_x) \left(\sum_{S_y \in \Omega} P(S_y|S_x, t_{xy})\, L_y \sum_{S_z \in \Omega} P(S_z|S_x, t_{xz})\, L_z \right)$$

where we are computing the likelihood of the subtree rooted at node x with direct descendants y and z, S_i denotes the character state of the i-th node, t_{ij} is the branch length (evolutionary time) between nodes i and j, and Ω is the set of all possible character states (for example, the nucleotides A, C, G, and T). Thus, the objective of ancestral reconstruction is to find the assignment to S_x for all x internal nodes that maximizes the likelihood of the observed data for a given tree.

Marginal and joint likelihood

Rather than compute the overall likelihood for alternative trees, the problem for ancestral reconstruction is to find the combination of character states at each ancestral node with the highest marginal ML. Generally speaking, there are two approaches to this problem. First, one may work upwards from the descendants of a tree to progressively assign the most likely character state to each ancestor taking into consideration only its immediate descendants. This approach is referred to as marginal reconstruction. It is akin to a greedy algorithm that makes the locally optimal choice at each stage of the optimization problem. While it can be highly efficient, it is not guaranteed to attain a globally optimal solution to the problem. Second, one may instead attempt to find the joint combination of ancestral character states throughout the tree that jointly maximizes the likelihood of the data. Thus, this approach is referred to as joint reconstruction. While it is not as rapid as marginal reconstruction, it is also less likely to be caught in the local optima in nonconvex objective functions that modern optimization methods and heuristics are designed to avoid. In the context of ancestral reconstruction, this means that a marginal reconstruction may assign a character state to the immediate ancestor that is locally optimal but deflects the joint distribution of ancestral character states away from the global optimum (for examples, see Pupko and colleagues [4]). Not surprisingly, joint reconstruction is more computationally complex than marginal reconstruction. Nevertheless, efficient algorithms for joint reconstruction have been developed with a time complexity that is generally linear with the number of observed taxa or sequences.

ML-based methods of ancestral reconstruction tend to provide greater accuracy than MP methods in the presence of variation in rates of evolution among characters (or across sites in a genome) [28,29]. However, these methods are not yet able to accommodate variation in rates of evolution over time, otherwise known as heterotachy. If the rate of evolution for a specific character accelerates on a branch of the phylogeny, then the amount of evolution that has occurred on that branch will be underestimated for a given length of the branch and assuming a constant rate of evolution for that character. In addition to that, it is difficult to distinguish heterotachy from variation among characters in rates of evolution [30].

Since ML (unlike maximum parsimony) requires the investigator to specify a model of evolution, its accuracy may be affected by the use of a grossly incorrect model (model misspecification). Furthermore, ML can only provide a single reconstruction of character states (what is often referred to as a "point estimate")—when the likelihood surface is highly nonconvex, comprising multiple peaks (local optima), then a single point estimate cannot provide an adequate representation, and a Bayesian approach may be more suitable.

Bayesian Inference

Bayesian inference uses the likelihood of observed data to update the investigator's belief, or prior distribution, to yield the posterior distribution. In the context of ancestral reconstruction, the objective is to infer the posterior probabilities of ancestral character states at each internal node of a given tree. Moreover, one can integrate these probabilities over the posterior distributions over the parameters of the evolutionary model and the space of all possible trees. This can be expressed as an application of Bayes' theorem:

$$P(S|D,\theta) = \frac{P(D|S,\theta)\,P(S|\theta)}{P(D|\theta)} \propto P(D|S,\theta)\,P(S|\theta)\,P(\theta),$$

where S represents the ancestral states, D corresponds to the observed data, and θ represents

both the evolutionary model and the phylogenetic tree. $P(D|S,\theta)$ is the likelihood of the observed data that can be computed by Felsenstein's pruning algorithm as given above. $P(S|\theta)$ is the prior probability of the ancestral states for a given model and tree. Finally, $P(D|\theta)$ is the probability of the data for a given model and tree, integrated over all possible ancestral states. We have given two formulations to emphasize the two different applications of Bayes' theorem, which we discuss in the following section.

Empirical and hierarchical Bayes

One of the first implementations of a Bayesian approach to ancestral sequence reconstruction was developed by Yang and colleagues, where the ML estimates of the evolutionary model and tree, respectively, were used to define the prior distributions. Thus, their approach is an example of an empirical Bayes method to compute the posterior probabilities of ancestral character states; this method was first implemented in the software package PAML [31]. In terms of the above Bayesian rule formulation, the empirical Bayes method fixes to the empirical estimates of the model and tree obtained from the data, effectively dropping from the posterior likelihood and prior terms of the formula. Moreover, Yang and colleagues [24] used the empirical distribution of site patterns (i.e., assignments of nucleotides to tips of the tree) in their alignment of observed nucleotide sequences in the denominator in place of exhaustively computing $P(D)$ over all possible values of S, given θ. Computationally, the empirical Bayes method is akin to the ML reconstruction of ancestral states except that, rather than searching for the ML assignment of states based on their respective probability distributions at each internal node, the probability distributions themselves are reported directly.

Empirical Bayes methods for ancestral reconstruction require the investigator to assume that the evolutionary model parameters and tree are known without error. When the size or complexity of the data makes this an unrealistic assumption, it may be more prudent to adopt the fully hierarchical Bayesian approach and infer the joint posterior distribution over the ancestral character states, model, and tree [32]. Huelsenbeck and Bollback [32] first proposed a hierarchical Bayes method to ancestral reconstruction by using Markov chain Monte Carlo (MCMC) methods to sample ancestral sequences from this joint posterior distribution. A similar approach was also used to reconstruct the evolution of symbiosis with algae in fungal species (lichenization) [33]. For example, the Metropolis-Hastings algorithm for MCMC explores the joint posterior distribution by accepting or rejecting parameter assignments on the basis of the ratio of posterior probabilities.

Put simply, the empirical Bayes approach calculates the probabilities of various ancestral states for a specific tree and model of evolution. By expressing the reconstruction of ancestral states as a set of probabilities, one can directly quantify the uncertainty for assigning any particular state to an ancestor. On the other hand, the hierarchical Bayes approach averages these probabilities over all possible trees and models of evolution, in proportion to how likely these trees and models are, given the data that has been observed.

However, whether the hierarchical Bayes method confers a substantial advantage in practice remains controversial [34]. Moreover, this fully Bayesian approach is limited to analyzing relatively small numbers of sequences or taxa because the space of all possible trees rapidly becomes too vast, making it computationally infeasible for chain samples to converge in a reasonable amount of time.

Calibration

Ancestral reconstruction can be informed by the observed states in historical samples of known age, such as fossils or archival specimens. Since the accuracy of ancestral reconstruction

generally decays with increasing time, the use of such specimens provides data that are closer to the ancestors being reconstructed and will most likely improve the analysis, especially when rates of character change vary through time. This concept has been validated by an experimental evolutionary study in which replicate populations of bacteriophage T7 were propagated to generate an artificial phylogeny [35]. In revisiting these experimental data, Oakley and Cunningham [36] found that maximum parsimony methods were unable to accurately reconstruct the known ancestral state of a continuous character (plaque size); these results were verified by computer simulation. This failure of ancestral reconstruction was attributed to a directional bias in the evolution of plaque size (from large to small plaque diameters) that required the inclusion of "fossilized" samples to address.

Studies of both mammalian carnivores [37] and fishes [38] have demonstrated that, without incorporating fossil data, the reconstructed estimates of ancestral body sizes are unrealistically large. Moreover, Graham Slater and colleagues showed using *Caniform carnivorans* that incorporating fossil data into prior distributions improved both the Bayesian inference of ancestral states and evolutionary model selection, relative to analyses using only contemporaneous data [39].

Models

Many models have been developed to estimate ancestral states of discrete and continuous characters from extant descendants [40]. Such models assume that the evolution of a trait through time may be modelled as a stochastic process. For discrete-valued traits (such as "pollinator type"), this process is typically taken to be a Markov chain; for continuous-valued traits (such as "brain size"), the process is frequently taken to be a Brownian motion or an Ornstein–Uhlenbeck process. Using this model as the basis for statistical inference, one can now use ML methods or Bayesian inference to estimate the ancestral states.

Discrete-state models

Suppose the trait in question may fall into one of k states, labeled $1,\ldots,k$. The typical means of modelling evolution of this trait is via a continuous-time Markov chain, which may be briefly described as follows (cf. Fig 2). Each state has associated to it rates of transition to all of the other states. The trait is modelled as stepping between the k states; when it reaches a given state, it starts an exponential "clock" for each of the other states that it can step to. It then "races" the clocks against each other, and it takes a step towards the state whose clock is the first to ring. In such a model, the parameters are the transition rates $\boldsymbol{q} = \{q_{ij}: 1 \leq i,j \leq k, i \neq j\}$, which can be estimated using, for example, ML methods, where one maximizes over the set of all possible configurations of states of the ancestral nodes.

In order to recover the state of a given ancestral node in the phylogeny (call this node α) by ML, the procedure is: find the ML estimate $\hat{\boldsymbol{q}}$ of $\boldsymbol{q}$; then compute the likelihood of each possible

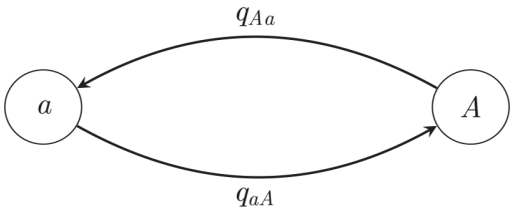


Fig 2. A general two-state Markov chain representing the rate of jumps from allele *a* to allele *A*. The different types of jumps are allowed to have different rates.

doi:10.1371/journal.pcbi.1004763.g002

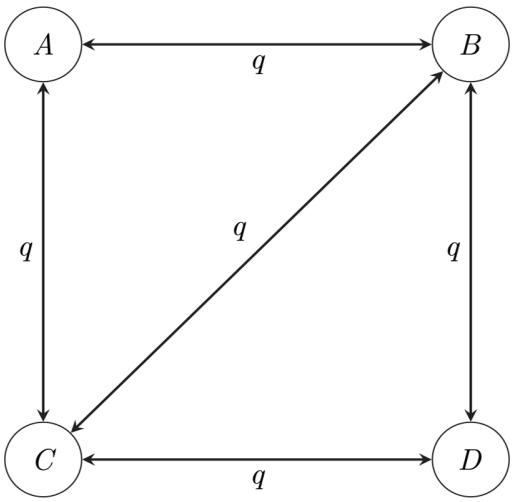


Fig 3. Example of a four-state 1 parameter Markov chain model. Note that in this diagram, transitions between states *A* and *D* have been disallowed; it is conventional to not draw the arrow rather than to draw it with a rate of 0.

doi:10.1371/journal.pcbi.1004763.g003

state for α conditioning on $\boldsymbol{q} = \hat{\boldsymbol{q}}$; finally, choose the ancestral state that maximizes this [19]. One may also use this substitution model as the basis for a Bayesian inference procedure, which would consider the posterior belief in the state of an ancestral node given some user-chosen prior.

Because such models may have as many as $k(k-1)$ parameters, overfitting may be an issue. Some common choices that reduce the parameter space are:

- Markov k-state 1 parameter model (Fig 3): this model is the reverse-in-time k-state counterpart of the Jukes–Cantor model. In this model, all transitions have the same rate q, regardless of their start and end states. Some transitions may be disallowed by declaring that their rates are simply 0; this may be the case, for example, if certain states cannot be reached from other states in a single transition.
- Asymmetrical Markov k-state 2 parameter model (Fig 4): in this model, the state space is ordered (so that, for example, state 1 is smaller than state 2, which is smaller than state 3, and transitions may only occur between adjacent states). This model contains two parameters q_{inc} and q_{dec}: one for the rate of increase of state (e.g., 0 to 1, 1 to 2, etc.) and one for the rate of decrease in state (e.g., from 2 to 1, 1 to 0, etc.).

Example: Binary state speciation and extinction model

The binary state speciation and extinction model [41] (BiSSE) is a discrete-space model that does not directly follow the framework of those mentioned above. It allows estimation of

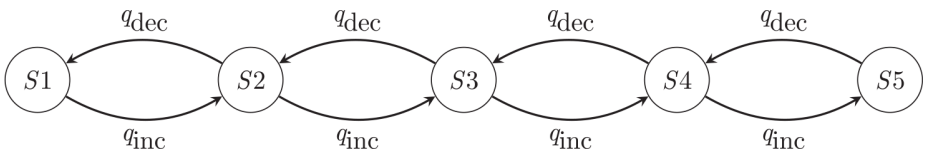


Fig 4. Graphical representation of an asymmetrical five-state 2-parameter Markov chain model.

doi:10.1371/journal.pcbi.1004763.g004

ancestral binary character states jointly with diversification rates associated with different character states; it may also be straightforwardly extended to a more general multiple-discrete-state model. In its most basic form, this model involves six parameters: two speciation rates (one each for lineages in states 0 and 1); similarly, two extinction rates; and two rates of character change. This model allows for hypothesis testing on the rates of speciation/extinction/character change, at the cost of increasing the number of parameters.

Continuous-state models

In the case where the trait instead takes nondiscrete values, one must instead turn to a model where the trait evolves as some continuous process. Inference of ancestral states by ML (or by Bayesian methods) would proceed as above but with the likelihoods of transitions in state between adjacent nodes given by some other continuous probability distribution.

- Brownian motion: in this case, if nodes U and V are adjacent in the phylogeny (say U is the ancestor of V) and separated by a branch of length t, the likelihood of a transition from U being in state x to V being in state y is given by a Gaussian density with mean 0 and variance $\sigma^2 t$. In this case, there is only one parameter (σ^2), and the model assumes that the trait evolves freely without a bias toward increase or decrease, and that the rate of change is constant throughout the branches of the phylogenetic tree [42].
- Ornstein–Uhlenbeck process: in brief, an Ornstein–Uhlenbeck process is a continuous stochastic process that behaves like a Brownian motion, but attracted toward some central value, where the strength of the attraction increases with the distance from that value. This is useful for modelling scenarios where the trait is subject to stabilizing selection around a certain value (say 0). Under this model, the above-described transition of U being in state x to V being in state y would have a likelihood defined by the transition density of an Ornstein–Uhlenbeck process with two parameters: σ^2, which describes the variance of the driving Brownian motion, and α, which describes the strength of its attraction to 0. As α tends to 0, the process is less and less constrained by its attraction to 0 and the process becomes a Brownian motion (Fig 5). Because of this, the models may be nested, and log-likelihood ratio tests discerning which of the two models is appropriate may be carried out.

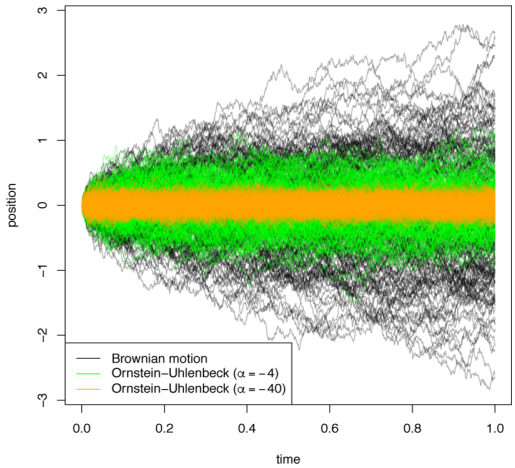


Fig 5. Plots of 200 trajectories of each of: Brownian motion with drift 0 and $\sigma^2 = 1$ (black); Ornstein–Uhlenbeck with $\sigma^2 = 1$ and $\alpha = -4$ (green); and Ornstein–Uhlenbeck with $\sigma^2 = 1$ and $\alpha = -40$ (orange).

doi:10.1371/journal.pcbi.1004763.g005

- Stable models of continuous character evolution [43]: though Brownian motion is appealing and tractable as a model of continuous evolution, it does not permit non-neutrality in its basic form, nor does it provide for any variation in the rate of evolution over time. Instead, one may use a stable process, one whose values at fixed times are distributed as stable distributions, to model the evolution of traits. Stable processes, roughly speaking, behave as Brownian motions that also incorporate discontinuous jumps. This allows us to appropriately model scenarios in which short bursts of fast trait evolution are expected. In this setting, ML methods are poorly suited because of a rugged likelihood surface and because the likelihood may be made arbitrarily large, so Bayesian methods are more appropriate.

Applications

Character evolution

Behaviour and life history evolution. *Diet reconstruction in Galapagos finches*: Both phylogenetic and character data are available for the radiation of finches inhabiting the Galapagos Islands. These data allow testing of hypotheses concerning the timing and ordering of character state changes through time via ancestral state reconstruction. During the dry season, the diets of the 13 species of Galapagos finches may be assorted into three broad diet categories, first those that consume grain-like foods are considered "granivores," those that ingest arthropods are termed "insectivores," and those that consume vegetation are classified as "folivores" [19]. Dietary ancestral state reconstruction under maximum parsimony recovers two major shifts from an insectivorous state: one to granivory and one to folivory. Maximum-likelihood ancestral state reconstruction recovers broadly similar results, with one significant difference: the common ancestor of the tree finch (*Camarhynchus*) and ground finch (*Geospiza*) clades is most likely granivorous rather than insectivorous (as judged by parsimony). In this case, this difference between ancestral states returned by maximum parsimony and ML likely occurs as a result of the fact that ML estimates consider branch lengths of the phylogenetic tree [19].

Morphological character evolution mammalian body mass. In an analysis of the body mass of 1,679 placental mammal species comparing stable models of continuous character evolution to Brownian motion models, Elliot and Mooers [43] showed that the evolutionary process describing mammalian body mass evolution is best characterized by a stable model of continuous character evolution, which accommodates rare changes of large magnitude. Under a stable model, ancestral mammals retained a low body mass through early diversification, with large increases in body mass coincident with the origin of several orders of large body massed species (e.g., ungulates). By contrast, simulation under a Brownian motion model recovered a less realistic, order of magnitude larger body mass among ancestral mammals, requiring significant reductions in body size prior to the evolution of orders exhibiting small body size (e.g., Rodentia). Thus, stable models recover a more realistic picture of mammalian body mass evolution by permitting large transformations to occur on a small subset of branches [43].

Correlated character evolution. Comparative methods (inferences drawn through comparison of related taxa) are often used to identify biological characteristics that do not evolve independently, which can reveal an underlying dependence. For example, the evolution of the shape of a finch's beak may be associated with its foraging behaviour. However, it is not advisable to search for these associations by the direct comparison of measurements or genetic sequences, as these observations are not independent because of their descent from common ancestors. For discrete characters, this problem was first addressed in the framework of

maximum parsimony by evaluating whether two characters tended to undergo a change on the same branches of the tree [44,45]. Felsenstein identified this problem for continuous character evolution and proposed a solution similar to ancestral reconstruction, in which the phylogenetic structure of the data was accommodated by directing the analysis on "independent contrasts" between nodes of the tree related by nonoverlapping branches [23].

Molecular evolution. On a molecular level, amino acid residues at different locations of a protein may evolve nonindependently, because they have a direct physicochemical interaction, or indirectly by their interactions with a common substrate, or through long-range interactions in the protein structure. Conversely, the folded structure of a protein could potentially be inferred from the distribution of residue interactions [46]. One of the earliest applications of ancestral reconstruction, to predict the three-dimensional structure of a protein through residue contacts, was published by Shindyalov and colleagues [47]. Phylogenies relating 67 different protein families were generated by a distance-based clustering method (unweighted pair group method with arithmetic mean, UPGMA), and ancestral sequences were reconstructed by parsimony. The authors reported a weak but significant tendency for coevolving pairs of residues to be colocated in the known three-dimensional structure of the proteins.

More recently, this concept has been applied to identify coevolving residues in protein sequences using more advanced methods for the reconstruction of phylogenies and ancestral sequences. For example, ancestral reconstruction has been used to identify coevolving residues in proteins encoded by RNA virus genomes, particularly in the human immunodeficiency virus (HIV) [48–50].

Vaccine design. RNA viruses such as HIV evolve at an extremely rapid rate, orders of magnitude faster than mammals or birds. For these organisms, ancestral reconstruction can be applied on a much shorter time scale; for example, in order to reconstruct the global or regional progenitor of an epidemic that has spanned decades rather than millions of years. A team around Brian Gaschen [51] proposed that such reconstructed strains be used as targets for vaccine design efforts as opposed to sequences isolated from patients in the present day. Because HIV is extremely diverse, a vaccine designed to work on one patient's viral population might not work for a different patient, because the evolutionary distance between these two viruses may be large. However, their most recent common ancestor is closer to each of the two viruses than they are to each other. Thus, a vaccine designed for a common ancestor could have a better chance of being effective for a larger proportion of circulating strains. Another team took this idea further by developing a center-of-tree (COT) reconstruction method to produce a sequence whose total evolutionary distance to contemporary strains is as small as possible [52]. Strictly speaking, this method was not ancestral reconstruction, as the COT sequence does not necessarily represent a sequence that has ever existed in the evolutionary history of the virus. However, Rolland and colleagues did find that, in the case of HIV, the COT virus was functional when synthesized. Similar experiments with synthetic ancestral sequences obtained by ML reconstruction have likewise shown that these ancestors are both functional and immunogenic, lending some credibility to these methods [53,54]. Furthermore, ancestral reconstruction can potentially be used to infer the genetic sequence of the transmitted HIV variants that have gone on to establish the next infection, with the objective of identifying distinguishing characteristics of these variants (as a nonrandom selection of the transmitted population of viruses) that may be targeted for vaccine design [55].

Genome rearrangements. Rather than inferring the ancestral DNA sequence, one may be interested in the larger-scale molecular structure of an ancestral genome. This problem is often approached in a combinatorial framework, by modelling genomes as permutations of genes or homologous regions. Various operations are allowed on these permutations, such as an inversion (a segment of the permutation is reversed in-place), deletion) (a segment is removed), or

transposition (a segment is removed from one part of the permutation and spliced in somewhere else). The "genome rearrangement problem," first posed by Watterson and colleagues [13], asks: given two genomes (permutations) and a set of allowable operations, what is the shortest sequence of operations that will transform one genome into the other? A generalization of this problem applicable to ancestral reconstruction is the "multiple genome rearrangement problem" [56]: given a set of genomes and a set of allowable operations, find (i) a binary tree with the given genomes as its leaves, and (ii) an assignment of genomes to the internal nodes of the tree, such that the total number of operations across the whole tree is minimized. This approach is similar to parsimony, except that the tree is inferred along with the ancestral sequences. Unfortunately, even the single genome rearrangement problem is NP-hard [57], although it has received much attention in mathematics and computer science (for a review, see Fertin and colleagues [58]).

Spatial applications migration. Ancestral reconstruction is not limited to biological traits. Spatial location is also a trait, and ancestral reconstruction methods can infer the locations of ancestors of the individuals under consideration. Such techniques were used by Lemey and colleagues to geographically trace the ancestors of 192 Avian influenza A-H5N1 strains sampled from 20 localities in Europe and Asia and of 101 rabies virus sequences sampled across 12 African countries [12].

Treating locations as discrete states (countries, cities, etc.) allows for the application of the discrete-state models described above. However, unlike in a model where the state space for the trait is small, there may be many locations, and transitions between certain pairs of states may rarely or never occur; for example, migration between distant locales may never happen directly if air travel between the two places does not exist, so such migrations must pass through intermediate locales first. This means that there could be many parameters in the model which are zero or close to zero. To this end, Lemey and colleagues used a Bayesian procedure to not only estimate the parameters and ancestral states but also to select which migration parameters are not zero; their work suggests that this procedure does lead to more efficient use of the data. They also explore the use of prior distributions that incorporate geographical structure or hypotheses about migration dynamics, finding that those they considered had little effect on the findings.

Using this analysis, the team around Lemey found that the most likely hub of diffusion of A-H5N1 is Guangdong, with Hong Kong also receiving posterior support. Further, their results support the hypothesis of long-standing presence of African rabies in West Africa [12].

Species ranges. Inferring historical biogeographic patterns often requires reconstructing ancestral ranges of species on phylogenetic trees [59]. For instance, a well-resolved phylogeny of plant species in the genus *Cyrtandra* [59] was used together with information of their geographic ranges to compare four methods of ancestral range reconstruction. The team compared Fitch parsimony [16] (FP; parsimony), stochastic mapping (SM; ML) [60], dispersal-vicariance analysis [61] (DIVA; parsimony), and dispersal-extinction-cladogenesis (DEC; maximum-likelihood) [11,62]. Results indicated that both parsimony methods performed poorly, which was likely due to the fact that parsimony methods do not consider branch lengths. Both maximum-likelihood methods performed better; however, DEC analyses that additionally allow incorporation of geological priors gave more realistic inferences about range evolution in *Cyrtandra* relative to other methods [59].

Another ML method recovers the phylogeographic history of a gene by reconstructing the ancestral locations of the sampled taxa [63]. This method assumes a spatially explicit random walk model of migration to reconstruct ancestral locations given the geographic coordinates of the individuals represented by the tips of the phylogenetic tree. When applied to a phylogenetic tree of chorus frogs *Pseudacris feriarum*, this method recovered recent northward expansion,

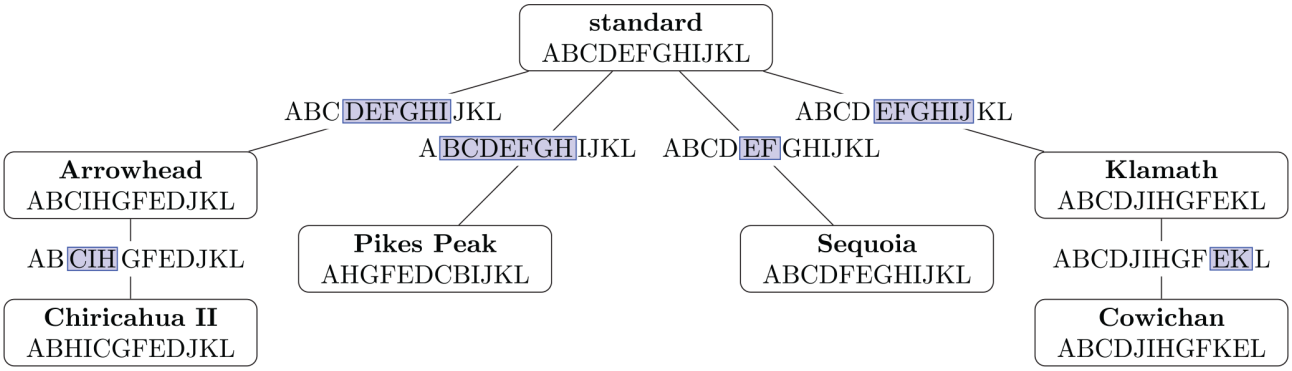


Fig 6. Phylogeny of seven regional strains of *Drosophila pseudoobscura*, as inferred by Sturtevant and Dobzhansky [64]. Displayed sequences do not correspond to the original paper, but were derived from the notation in the authors' companion paper [8] as follows: A (63A–65B), B (65C–68D), C (69A–70A), D (70B–70D), E (71A–71B), F (71A–73C), G (74A–74C), H (75A–75C), I (76A–76B), J (76C–77B), K (78A–79D), L (80A–81D). Inversions inferred by the authors are highlighted in blue along branches.

doi:10.1371/journal.pcbi.1004763.g006

higher per-generation dispersal distance in the recently colonized region, a noncentral ancestral location, and directional migration.

The first consideration of the multiple genome rearrangement problem, long before its formalization in terms of permutations, was presented by Sturtevant and Dobzhansky in 1936 (Fig 6) [64]. They examined genomes of several strains of fruit fly from different geographic locations, and observed that one configuration, which they called "standard," was the most common throughout all the studied areas. Remarkably, they also noticed that four different strains could be obtained from the standard sequence by a single inversion, and two others could be related by a second inversion. This allowed them to hypothesize a phylogeny for the sequences and to infer that the standard sequence was probably also the ancestral one.

Linguistic evolution. Reconstructions of the words and phenomes of ancient protolanguages such as Proto-Indo-European have been performed based on the observed analogues in present-day languages. Typically, these analyses are carried out manually using the "comparative method" [65]. First, words from different languages with a common etymology (cognates) are identified in the contemporary languages under study, analogous to the identification of orthologous biological sequences. Second, correspondences between individual sounds in the cognates are identified, a step similar to biological sequence alignment, although performed manually. Finally, likely ancestral sounds are hypothesized by manual inspection and various heuristics (such as the fact that most languages have both nasal and non-nasal vowels) [65].

Software

There are many software packages available that perform ancestral state reconstruction. Generally, these software packages have been developed and maintained through the efforts of scientists in related fields and released under free software licenses. The following table is not meant to be a comprehensive itemization of all available packages, but provides a representative sample of the extensive variety of packages that implement methods of ancestral reconstruction with different strengths and features (Table 1).

Package Descriptions

Molecular evolution

The majority of these software packages are designed for analyzing genetic sequence data. For example, PAML [31] is a collection of programs for the phylogenetic analysis of DNA and

Table 1. List of software for ancestral reconstruction.

Name	Methods	Platform	Supported Input Formats	Character Types	Continuous (C) or Discrete (D) Characters	Software Licence
PAML	ML	Unix, Mac, Win	PHYLIP, NEXUS, FASTA	Nucleotide, Protein	D	Proprietary
BEAST2	Bayesian	Unix, Mac, Win	NEXUS, BEAST XML	Nucleotide, Protein, Geographic	C, D	GNU Lesser General Public License
APE	ML	Unix, Mac, Win	NEXUS, FASTA, CLUSTAL	Nucleotide, Protein	C, D	GNU General Public License
Diversitree	ML	Unix, Mac, Win	NEXUS	Qualitative and quantitative traits, Geographic	C, D	GNU General Public License, version 2
HyPhy	ML	Unix, Mac, Win	MEGA, NEXUS, FASTA, PHYLIP	Nucleotide, Protein (customizable)	D	GNU Free Documentation License 1.3
BayesTraits	Bayesian	Unix, Mac, Win	TSV or space delimited table. Rows are species, columns are traits.	Qualitative and quantitative traits	C, D	Creative Commons Attribution License
Lagrange	ML	Linux, Mac, Win	TSV/CSV of species regions. Rows are species and columns are geographic regions	Geographic	-	GNU General Public License, version 2
Mesquite	Parsimony, ML	Unix, Mac, Win	Fasta, NBRF, Genbank, PHYLIP, CLUSTAL, TSV	Nucleotide, Protein, Geographic	C, D	Creative Commons Attribution 3.0 License
Phylomapper	ML, Bayesian (version 2)	Unix, Mac, Win	NEXUS	Geographic, Ecological niche	C, D	-
Ancestors	ML	Web	Fasta	Nucleotide (indels)	D	-
Phyrex	Maximum Parsimony	Linux	Fasta	Gene expression	C, D	Proprietary
SIMMAP	SM	Mac	XML-like format	Nucleotide, qualitative traits	D	Proprietary
MrBayes	Bayesian	Unix, Mac, Win	NEXUS	Nucleotide, Protein	D	GNU General Public License
PARANA	Maximum Parsimony	Unix, Mac, Win	Newick	Biological networks	D	Apache License
PHAST (PREQUEL)	ML	Unix, Mac, Win	Multiple Alignment	Nucleotide	D	BSD License
RASP	ML, Bayesian	Unix, Mac, Win	Newick	Geographic	D	-
VIP	Maximum Parsimony	Linux, Win	Newick	Geographic	D (grid)	GPL Creative Commons
FastML	ML	Web, Unix	Fasta	Nucleotide, Protein	D	Copyright
MLGO	ML	Web	Custom	Gene order permutation	D	GNU
BADGER	Bayesian	Unix, Mac, Win	Custom	Gene order permutation	D	GNU GPL version 2
COUNT	Maximum Parsimony	Unix, Mac, Win	Tab-delimited text file of rows for taxa and count data in columns	Count data (homolog family size)	D	BSD
MEGA	Maximum parsimony, ML	Mac, Win	MEGA	Nucleotide, Protein	D	Proprietary
ANGES	Local Parsimony	Unix	Custom	Genome maps	D	GNU General Public License, version 3
EREM	ML	Win, Unix, Matlab module	Custom text format for model parameters, tree, observed character values	Binary	D	None specified, although site indicates software is freely available

doi:10.1371/journal.pcbi.1004763.t001

protein sequence alignments by ML. Ancestral reconstruction can be performed using the codeml program. HyPhy, Mesquite, and MEGA are also software packages for the phylogenetic analysis of sequence data, but are designed to be more modular and customizable. HyPhy [66] implements a joint ML method of ancestral sequence reconstruction [4] that can be readily adapted to reconstructing a more generalized range of discrete ancestral character states such as geographic locations by specifying a customized model in its batch language. Mesquite [67] provides ancestral state reconstruction methods for both discrete and continuous characters using both maximum parsimony and ML methods. It also provides several visualization tools for interpreting the results of ancestral reconstruction. MEGA [68] is a modular system, too, but places greater emphasis on ease-of-use than customization of analyses. As of version 5, MEGA allows the user to reconstruct ancestral states using maximum parsimony, ML, and empirical Bayes methods [68].

The Bayesian analysis of genetic sequences may confer greater robustness to model misspecification. MrBayes [69] allows inference of ancestral states at ancestral nodes using the full hierarchical Bayesian approach. The PREQUEL program distributed in the PHAST package performs comparative evolutionary genomics using ancestral sequence reconstruction [70]. SIMMAP stochastically maps mutations on phylogenies [71]. BayesTraits [26] analyses discrete or continuous characters in a Bayesian framework to evaluate models of evolution, reconstruct ancestral states, and detect correlated evolution between pairs of traits.

Other character types. Other software packages are more oriented towards the analysis of qualitative and quantitative traits (phenotypes). For example, the ape package [72] in the statistical computing environment R also provides methods for ancestral state reconstruction for both discrete and continuous characters through the *ace* function, including ML. Note that *ace* performs reconstruction by computing scaled conditional likelihoods instead of the marginal or joint likelihoods used by other ML-based methods for ancestral reconstruction, which may adversely affect the accuracy of reconstruction at nodes other than the root. Phyrex implements a maximum parsimony-based algorithm to reconstruct ancestral gene expression profiles in addition to a ML method for reconstructing ancestral genetic sequences (by wrapping around the baseml function in PAML) [73].

Several software packages also reconstruct phylogeography. BEAST (Bayesian Evolutionary Analysis by Sampling Trees [74]) provides tools for reconstructing ancestral geographic locations from observed sequences annotated with location data using Bayesian MCMC sampling methods. Diversitree [75] is an R package providing methods for ancestral state reconstruction under Mk2 (a continuous time Markov model of binary character evolution [76]) and BiSSE models. Lagrange performs analyses on reconstruction of geographic range evolution on phylogenetic trees [11]. Phylomapper [63] is a statistical framework for estimating historical patterns of gene flow and ancestral geographic locations. RASP [77] infers ancestral state using statistical DIVA, Lagrange, Bayes-Lagrange, BayArea, and BBM methods. VIP [78] infers historical biogeography by examining disjunct geographic distributions.

Genome rearrangements provide valuable information in comparative genomics between species. ANGES [79] compares extant-related genomes through ancestral reconstruction of genetic markers. BADGER [80] uses a Bayesian approach to examining the history of gene rearrangement. Count [81] reconstructs the evolution of the size of gene families. EREM [82] analyses the gain and loss of genetic features encoded by binary characters. PARANA [83] performs parsimony-based inference of ancestral biological networks that represent gene loss and duplication.

Web applications. Finally, there are several web server-based applications that allow investigators to use ML methods for ancestral reconstruction of different character types without having to install any software. For example, Ancestors [84] is a web server for ancestral

genome reconstruction by the identification and arrangement of syntenic regions. FastML [85] is a web server for probabilistic reconstruction of ancestral sequences by ML that uses a gap character model for reconstructing indel variation. MLGO [86] is a web server for ML gene order analysis.

Future Directions

The development and application of computational algorithms for ancestral reconstruction continues to be an active area of research across disciplines. For example, the reconstruction of sequence insertions and deletions (indels) has lagged behind the more straightforward application of substitution models. Bouchard-Côté and Jordan recently described a new model (the Poisson Indel Process [87]) that represents an important advance on the archetypal Thorne-Kishino-Felsenstein model of indel evolution [88]. In addition, the field is being driven forward by rapid advances in the area of next-generation sequencing technology, where sequences are generated from millions of nucleic acid templates by extensive parallelization of sequencing reactions in a custom apparatus. These advances have made it possible to generate a "deep" snapshot of the genetic composition of a rapidly-evolving population, such as RNA viruses [89] or tumour cells [90], in a relatively short amount of time. At the same time, the massive amount of data and platform-specific sequencing error profiles has created new bioinformatic challenges for processing these data for ancestral sequence reconstruction.

The version history of the text file and the peer reviews (and response to reviews) are available as supporting information in S1 and S2 Texts.

Supporting Information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews. Human-readable versions of the reviews and authors' responses are available as comments on this article.
(XML)

References

1. Platnick NI, Cameron HD. Cladistic methods in textual, linguistic, and phylogenetic analysis. Systematic Biology. 1977; 26(4):380–5.
2. Tehrani JJ. The phylogeny of little red riding hood. PloS ONE. 2013; 8(11):e78871. doi: 10.1371/journal.pone.0078871 PMID: 24236061
3. Walker RS, Hill KR, Flinn MV, Ellsworth RM. Evolutionary history of hunter-gatherer marriage practices. PLoS ONE. 2011; 6(4):e19066. doi: 10.1371/journal.pone.0019066 PMID: 21556360
4. Pupko T, Pe I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution. 2000; 17(6):890–6. PMID: 10833195
5. Pagel M, Meade A, Barker D. Bayesian Estimation of Ancestral Character States on Phylogenies. Systematic Biology. 2004; 53(5):673–84. PMID: 15545248
6. Sanger F, Thompson EOP, Kitai R. The amide groups of insulin. Biochemical Journal. 1955; 59(3):509. PMID: 14363129
7. Pauling L, Zuckerkand E. Chemical Paleogenetics: Molecular "Restoration Studies" of Extinct Forms of Life. Acta Chemica Scandinavica. 1963; 17:9–16. doi: 10.3891/acta.chem.scand.17s-0009
8. Dobzhansky T, Sturtevant AH. Inversions in the chromosomes of Drosophila pseudoobscura. Genetics. 1938; 23(1):28. PMID: 17246876
9. Harms MJ, Thornton JW. Analyzing protein structure and function using ancestral gene reconstruction. Current opinion in structural biology. 2010; 20(3):360–6. doi: 10.1016/j.sbi.2010.03.005 PMID: 20413295

10. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2006; 2(6):e69. PMID: 16789817
11. Ree RH, Smith SA. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology. 2008; 57(1):4–14. doi: 10.1080/10635150701883881 PMID: 18253896
12. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 2009; 5.
13. Watterson GA, Ewens WJ, Hall TE, Morgan A. The chromosome inversion problem. Journal of Theoretical Biology. 1982; 99(1):1–7.
14. Tuffley C, Steel M. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of mathematical biology. 1997; 59(3):581–607. PMID: 9172826
15. Swofford DL, Maddison WP. Reconstructing ancestral character states under Wagner parsimony. Mathematical Biosciences. 1987; 87(2):199–229.
16. Fitch WM. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology. 1971; 20(4):406–16.
17. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–90. PMID: 16928733
18. Sankoff D. Minimal mutation trees of sequences. SIAM Journal on Applied Mathematics. 1975; 28 (1):35–42.
19. Schluter D, Price T, Mooers AO, Ludwig D. Likelihood of ancestor states in adaptive radiation. Evolution. 1997; 51(6):1699–711. PMID: ISI:000071652200001.
20. Felsenstein J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology. 1973; 22(3):240–9.
21. Cunningham CW, Omland KE, Oakley TH. Reconstructing ancestral character states: a critical reappraisal. Trends in Ecology & Evolution. 1998; 13(9):361–6.
22. Li G, Steel M, Zhang L. More taxa are not necessarily better for the reconstruction of ancestral character states. Systematic biology. 2008; 57(4):647–53. doi: 10.1080/10635150802203898 PMID: 18709600
23. Felsenstein J. Phylogenies and the comparative method. American Naturalist. 1985:1–15.
24. Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics. 1995; 141(4):1641–50. PMID: 8601501
25. Koshi JM, Goldstein RA. Probabilistic reconstruction of ancestral protein sequences. Journal of Molecular Evolution. 1996; 42(2):313–20. PMID: 8919883
26. Pagel M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic biology. 1999; 48(3):612–22.
27. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981; 17(6):368–76. PMID: 7288891
28. Eyre-Walker A. Problems with parsimony in sequences of biased base composition. Journal of Molecular Evolution. 1998; 47(6):686–90. PMID: 9847410
29. Pupko T, Pe'er I, Hasegawa M, Graur D, Friedman N. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. Bioinformatics. 2002; 18(8):1116–23. PMID: 12176835
30. Gruenheit N, Lockhart PJ, Steel M, Martin W. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Molecular Biology and Evolution. 2008; 25(7):1512–20. doi: 10.1093/molbev/msn098 PMID: 18424773
31. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007; 24(8):1586–91. PMID: 17483113
32. Huelsenbeck JP, Bollback JP. Empirical and hierarchical Bayesian estimation of ancestral states. Systematic Biology. 2001; 50(3):351–66. PMID: ISI:000169823200006.
33. Lutzoni F, Pagel M, Reeb V. Major fungal lineages are derived from lichen symbiotic ancestors. Nature. 2001; 411(6840):937–40. PMID: 11418855
34. Hanson-Smith V, Kolaczkowski B, Thornton JW. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Molecular biology and evolution. 2010; 27(9):1988–99. doi: 10.1093/molbev/msq081 PMID: 20368266
35. Hillis DM, Bull JJ, White ME, Badgett MR, Molineux IJ. Experimental phylogenetics: generation of a known phylogeny. Science. 1992; 255(5044):589–92. PMID: 1736360
36. Oakley TH, Cunningham CW. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. Evolution. 2000; 54(2):397–405. PMID: 10937216

37. Finarelli JA, Flynn JJ. Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): the effects of incorporating data from the fossil record. Systematic Biology. 2006; 55(2):301–13. PMID: 16611601

38. Albert JS, Johnson DM, Knouft JH. Fossils provide better estimates of ancestral body size than do extant taxa in fishes. Acta Zoologica. 2009; 90(s1):357–84.

39. Slater GJ, Harmon LJ, Alfaro ME. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution. 2012; 66(12):3931–44. doi: 10.1111/j.1558-5646.2012.01723.x PMID: 23206147

40. Webster AJ, Purvis A. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. Proceedings of the Royal Society of London B: Biological Sciences. 2002; 269 (1487):143–9.

41. Maddison WP, Midford PE, Otto SP. Estimating a binary character's effect on speciation and extinction. Systematic biology. 2007; 56(5):701–10. PMID: 17849325

42. Martins EP. Estimating the rate of phenotypic evolution from comparative data. American Naturalist. 1994:193–209.

43. Elliot MG, Mooers AØ. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC evolutionary biology. 2014; 14(1):226.

44. Ridley M. The explanation of organic diversity: the comparative method and adaptations for mating. Oxford: Clarendon Press; 1983.

45. Maddison WP. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution. 1990:539–57.

46. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics. 1994; 18(4):309–17.

47. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engineering. 1994; 7(3):349–58. PMID: 8177884

48. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proceedings of the National Academy of Sciences. 1993; 90(15):7176–80.

49. Shapiro B, Rambaut A, Pybus OG, Holmes EC. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Molecular biology and evolution. 2006; 23(9):1724–30. PMID: 16774976

50. Poon AFY, Lewis FI, Pond SLK, Frost SDW. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol. 2007; 3(11):e231. PMID: 18039027

51. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. Science. 2002; 296(5577):2354–60. PMID: 12089434

52. Rolland M, Jensen MA, Nickle DC, Yan J, Learn GH, Heath L, et al. Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. Journal of virology. 2007; 81 (16):8507–14. PMID: 17537854

53. Kothe DL, Li Y, Decker JM, Bibollet-Ruche F, Zammit KP, Salazar MG, et al. Ancestral and consensus envelope immunogens for HIV-1 subtype C. Virology. 2006; 352(2):438–49. PMID: 16780913

54. Doria-Rose NA, Learn GH, Rodrigo AG, Nickle DC, Li F, Mahalanabis M, et al. Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. Journal of virology. 2005; 79 (17):11214–24. PMID: 16103173

55. McCloskey RM, Liang RH, Harrigan PR, Brumme ZL, Poon AFY. An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. Journal of virology. 2014; 88(11):6181–94. doi: 10.1128/JVI.00483-14 PMID: 24648453

56. Bourque G, Pevzner PA. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome research. 2002; 12(1):26–36. PMID: 11779828

57. Even S, Goldreich O. The minimum-length generator sequence problem is NP-hard. Journal of Algorithms. 1981; 2(3):311–3.

58. Fertin G. Combinatorics of genome rearrangements: MIT press; 2009.

59. Clark JR, Ree RH, Alfaro ME, King MG, Wagner WL, Roalson EH. A comparative study in ancestral range reconstruction methods: retracing the uncertain histories of insular lineages. Systematic Biology. 2008; 57(5):693–707. doi: 10.1080/10635150802426473 PMID: 18853357

60. Huelsenbeck JP, Nielsen R, Bollback JP. Stochastic mapping of morphological characters. Systematic Biology. 2003; 52(2):131–58. PMID: ISI:000181839200001.

61. Ronquist F. DIVA version 1.2. Computer program for MacOS and Win32. Evolutionary Biology Centre, Uppsala University Available at http://www.ebcuuse/systzoo/research/diva/diva html. 2001.
62. Ree RH, Moore BR, Webb CO, Donoghue MJ. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. Evolution. 2005; 59(11):2299–311. PMID: 16396171
63. Lemmon AR, Lemmon EM. A likelihood framework for estimating phylogeographic history on a continuous landscape. Systematic Biology. 2008; 57(4):544–61. doi: 10.1080/10635150802304761 PMID: 18686193
64. Sturtevant AH, Dobzhansky T. Geographical Distribution and Cytology of" Sex Ratio" in Drosophila Pseudoobscura and Related Species. Genetics. 1936; 21(4):473. PMID: 17246805
65. Campbell L. Historical linguistics: Edinburgh University Press; 2013.
66. Pond SLK, Muse SV. HyPhy: hypothesis testing using phylogenies. Statistical methods in molecular evolution: Springer; 2005. p. 125–81.
67. Maddison W, Maddison D. Mesquite: a modular system for evolutionary analysis. 2.75 ed20011.
68. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular biology and evolution. 2007; 24(8):1596–9. PMID: 17488738
69. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17(8):754–5. PMID: ISI:000171021000016.
70. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in bioinformatics. 2011; 12(1):41–51. doi: 10.1093/bib/bbq072 PMID: 21278375
71. Bollback JP. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC bioinformatics. 2006; 7(1):88.
72. Paradis E. Analysis of phylogenetics and evolution with R. New York: Springer; 2006.
73. Rossnes R, Eidhammer I, Liberles DA. Phylogenetic reconstruction of ancestral character states for gene expression and mRNA splicing data. BMC bioinformatics. 2005; 6(1):127.
74. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C- H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014; 10(4):e1003537. doi: 10.1371/journal.pcbi.1003537 PMID: 24722319
75. FitzJohn RG. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution. 2012; 3(6):1084–92.
76. Pagel M. Detecting Correlated Evolution on Phylogenies—a General- Method for the Comparative-Analysis of Discrete Characters. Proceedings of the Royal Society of London Series B-Biological Sciences. 1994; 255(1342):37–45. PMID: ISI:A1994MT55500006.
77. Yu Y, Harris AJ, Blair C, He X. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Molecular Phylogenetics and Evolution. 2015; 87:46–9. doi: 10.1016/j.ympev.2015.03.008 PMID: 25819445
78. Arias JS, Szumik CA, Goloboff PA. Spatial analysis of vicariance: a method for using direct geographical information in historical biogeography. Cladistics. 2011; 27(6):617–28.
79. Jones BR, Rajaraman A, Tannier E, Chauve C. ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics. 2012; 28(18):2388–90. doi: 10.1093/bioinformatics/bts457 PMID: 22820205
80. Larget B, Kadane JB, Simon DL. A Bayesian approach to the estimation of ancestral genome arrangements. Molecular phylogenetics and evolution. 2005; 36(2):214–23. PMID: 15893477
81. Csűös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010; 26(15):1910–2. doi: 10.1093/bioinformatics/btq315 PMID: 20551134
82. Affre L, Thompson JD, Debussche M. Genetic structure of continental and island populations of the Mediterranean endemic Cyclamen balearicum (Primulaceae). American Journal of Botany. 1997; 84 (4):437–51. PMID: ISI:A1997WV34000002.
83. Patro R, Sefer E, Malin J, Marçais G, Navlakha S, Kingsford C. Parsimonious reconstruction of network evolution. Algorithms for Molecular Biology. 2012; 7(1):1.
84. Diallo AB, Makarenkov V, Blanchette M. Ancestors 1.0: a web server for ancestral sequence reconstruction. Bioinformatics. 2010; 26(1):130–1. doi: 10.1093/bioinformatics/btp600 PMID: 19850756
85. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic acids research. 2012; 40(W1): W580–W4.
86. Hu F, Lin Y, Tang J. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC bioinformatics. 2014; 15(1):1.
87. Bouchard-Côté A, Jordan MI. Evolutionary inference via the Poisson indel process. Proceedings of the National Academy of Sciences. 2013; 110(4):1160–6.

88. Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. Journal of Molecular Evolution. 1991; 33(2):114–24. PMID: 1920447

89. Poon AFY, Swenson LC, Bunnik EM, Edo-Matas D, Schuitemaker H, van't Wout AB, et al. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLoS Comput Biol. 2012; 8(11):e1002753. doi: 10.1371/journal.pcbi.1002753 PMID: 23133358

90. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014; 10(4):e1003535. doi: 10.1371/journal.pcbi.1003535 PMID: 24743184

D.4 Phylogeny

Further explanations for Chapter 4.

D.4.1 Basic definitions

- Jacques et al. (2023) — A step-by-step practical guide to gene and protein phylogenetics, from homologue collection and alignment to tree building and interpretation. Useful as a broad overview of the whole workflow behind the definitions in this section. *(tutorial; CC BY 4.0)*

  Roadmap to the study of gene and protein phylogeny and evolution—A practical guide. Jacques et al. (2023). *PLOS ONE*. doi:10.1371/journal.pone.0279597

  > Developments in sequencing technologies and the sequencing of an ever-increasing number of genomes have revolutionised studies of biodiversity and organismal evolution. This accumulation of data has been paralleled by the creation of numerous public biological databases through which the scientific community can mine the sequences and annotations of genomes, transcriptomes, and proteomes of multiple species. However, to find the appropriate databases and bioinformatic tools for respective inquiries and aims can be challenging. Here, we present a compilation of DNA and protein databases, as well as bioinformatic tools for phylogenetic reconstruction and a wide range of studies on molecular evolution. We provide a protocol for information extraction from biological databases and simple phylogenetic reconstruction using probabilistic and distance methods, facilitating the study of biodiversity and evolution at the molecular level for the broad scientific community.

- Gregory (2008) — An introductory article on how to read phylogenetic trees, covering nodes, rooting, clades and common misinterpretations. Helpful for the tree terminology subsection. *(primer; CC BY-NC 2.0 (link only))*

  Understanding Evolutionary Trees. Gregory (2008). *Evolution: Education and Outreach*. doi:10.1007/s12052-008-0035-x

  > Charles Darwin sketched his first evolutionary tree in 1837, and trees have remained a central metaphor in evolutionary biology up to the present. Today, phylogenetics—the science of constructing and evaluating hypotheses about historical patterns of descent in the form of evolutionary trees—has become pervasive within and increasingly outside evolutionary biology. Fostering skills in "tree thinking" is therefore a critical component of biological education. Conversely, misconceptions about evolutionary trees can be very detrimental to one's understanding of the patterns and processes that have occurred in the history of life. This paper provides a basic introduction to evolutionary trees, including some guidelines for how and how not to read them. Ten of the most common misconceptions about evolutionary trees and their implications for understanding evolution are addressed.

D.4.2 Phylogenetic methods

- Aris-Brosou and Rodrigue (2019) — An introduction to substitution models, likelihood and Bayesian inference in molecular evolution. Complements the sections on maximum likelihood, Bayesian inference and substitution models. *(book chapter; CC BY 4.0)*

  A Not-So-Long Introduction to Computational Molecular Evolution. Aris-Brosou & Rodrigue (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_3

  > In this chapter, we give a not-so-long and self-contained introduction to computational molecular evolution. In particular, we present the emergence of the use of likelihood-based methods, review the standard DNA substitution models, and introduce how model choice operates. We also present recent developments in inferring absolute divergence times and rates on a phylogeny, before showing how state-of-the-art models take inspiration from diffusion theory to link population genetics, which traditionally focuses at a taxonomic level below that of the species, and molecular evolution. Although this is not a cookbook chapter, we try and point to popular programs and implementations along the way.

- Kellis et al. (2021) — Lecture-note chapter from an MIT course that works through distance-based methods (UPGMA, neighbour joining) and character-based methods (parsimony, likelihood) with worked examples. *(book chapter; CC BY-NC-SA 4.0)*

  Molecular Evolution and Phylogenetics (Chapter 26 of Computational Biology: Genomes, Networks, and Evolution). Kellis et al. (2021). *LibreTexts (MIT open textbook)*. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/26%3A_Molecular_Evolution_and_Phylogenetics

- Reis and Yang (2019) — A tutorial on Bayesian divergence-time estimation with relaxed clocks and fossil calibrations in MCMCtree. Useful for the Bayesian inference subsection and for how trees are dated. *(book chapter; CC BY 4.0)*

  Bayesian Molecular Clock Dating Using Genome-Scale Datasets. dos Reis & Yang (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_10

  > Bayesian methods for molecular clock dating of species divergences have been greatly developed during the past decade. Advantages of the methods include the use of relaxed-clock models to describe evolutionary rate variation in the branches of a phylogenetic tree and the use of flexible fossil calibration densities to describe the uncertainty in node ages. The advent of next-generation sequencing technologies has led to a flood of genome-scale datasets for organisms belonging to all domains in the tree of life. Thus, a new era has begun where dating the tree of life using genome-scale data is now within reach. In this protocol, we explain how to use the computer program MCMCTree to perform Bayesian inference of divergence times using genome-scale datasets. We use a ten-species primate phylogeny, with a molecular alignment of over three million base pairs, as an exemplar on how to carry out the analysis. We pay particular attention to how to set up the analysis and the priors and how to diagnose the MCMC algorithm used to obtain the posterior estimates of divergence times and evolutionary rates.

D.4.3 Alignments, genomes and conflicting trees

- Bravo et al. (2019) — A broad review of phylogenomic practice, including data collection, gene-tree heterogeneity, incomplete lineage sorting and coalescent-based species-tree methods. Covers most of the topics of this section. *(review; CC BY 4.0)*

  Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. Bravo et al. (2019). *PeerJ*. doi:10.7717/peerj.6399

  > Building the Tree of Life (ToL) is a major challenge of modern biology, requiring advances in cyberinfrastructure, data collection, theory, and more. Here, we argue that phylogenomics stands to benefit by embracing the many heterogeneous genomic signals emerging from the first decade of large-scale phylogenetic analysis spawned by high-throughput sequencing (HTS). Such signals include those most commonly encountered in phylogenomic datasets, such as incomplete lineage sorting, but also those reticulate processes emerging with greater frequency, such as recombination and introgression. Here we focus specifically on how phylogenetic methods can accommodate the heterogeneity incurred by such population genetic processes; we do not discuss phylogenetic methods that ignore such processes, such as concatenation or supermatrix approaches or supertrees. We suggest that methods of data acquisition and the types of markers used in phylogenomics will remain restricted until a posteriori methods of marker choice are made possible with routine whole-genome sequencing of taxa of interest. We discuss limitations and potential extensions of a model supporting innovation in phylogenomics today, the multispecies coalescent model (MSC). Macroevolutionary models that use phylogenies, such as character mapping, often ignore the heterogeneity on which building phylogenies increasingly rely and suggest that assimilating such heterogeneity is an important goal moving forward. Finally, we argue that an integrative cyberinfrastructure linking all steps of the process of building the ToL, from specimen acquisition in the field to publication and tracking of phylogenomic data, as well as a culture that values contributors at each step, are essential for progress.

- Liu et al. (2019) — Explains the multispecies coalescent and the main summary and full-likelihood methods for estimating species trees from many gene trees. Useful background for the gene-tree/species-tree subsection. *(book chapter; CC BY 4.0)*

Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model. Liu et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_7

> The multispecies coalescent (MSC) model provides a compelling framework for building phylogenetic trees from multilocus DNA sequence data. The pure MSC is best thought of as a special case of so-called "multispecies network coalescent" models, in which gene flow is allowed among branches of the tree, whereas MSC methods assume there is no gene flow between diverging species. Early implementations of the MSC, such as "parsimony" or "democratic vote" approaches to combining information from multiple gene trees, as well as concatenation, in which DNA sequences from multiple gene trees are combined into a single "supergene," were quickly shown to be inconsistent in some regions of tree space, in so far as they converged on the incorrect species tree as more gene trees and sequence data were accumulated. The anomaly zone, a region of tree space in which the most frequent gene tree is different from the species tree, is one such region where many so-called "coalescent" methods are inconsistent. Second-generation implementations of the MSC employed Bayesian or likelihood models; these are consistent in all regions of gene tree space, but Bayesian methods in particular are incapable of handling the large phylogenomic data sets currently available. Two-step methods, such as MP-EST and ASTRAL, in which gene trees are first estimated and then combined to estimate an overarching species tree, are currently popular in part because they can handle large phylogenomic data sets. These methods are consistent in the anomaly zone but can sometimes provide inappropriate measures of tree support or apportion error and signal in the data inappropriately. MP-EST in particular employs a likelihood model which can be conveniently manipulated to perform statistical tests of competing species trees, incorporating the likelihood of the collected gene trees on each species tree in a likelihood ratio test. Such tests provide a useful alternative to the multilocus bootstrap, which only indirectly tests the appropriateness of competing species trees. We illustrate these tests and implementations of the MSC with examples and suggest that MSC methods are a useful class of models effectively using information from multiple loci to build phylogenetic trees.

- Bjornson et al. (2024) — Reviews hybridisation, introgression and horizontal transfer as sources of gene-tree conflict, and the methods (including phylogenetic networks) used to detect them. Relevant to the subsection on phylogenetic networks. *(review; CC BY 4.0)*

  Reticulate evolution: Detection and utility in the phylogenomics era. Bjornson et al. (2024). *Molecular Phylogenetics and Evolution*. doi:10.1016/j.ympev.2024.108197

  > Phylogenomics has enriched our understanding that the Tree of Life can have network-like or reticulate structures among some taxa and genes. Two non-vertical modes of evolution - hybridization/introgression and horizontal gene transfer - deviate from a strictly bifurcating tree model, causing non-treelike patterns. However, these reticulate processes can produce similar patterns to incomplete lineage sorting or recombination, potentially leading to ambiguity. Here, we present a brief overview of a phylogenomic workflow for inferring organismal histories and compare methods for distinguishing modes of reticulate evolution. We discuss how the timing of coalescent events can help disentangle introgression from incomplete lineage sorting and how horizontal gene transfer events can help determine the relative timing of speciation events. In doing so, we identify pitfalls of certain methods and discuss how to extend their utility across the Tree of Life. Workflows, methods, and future directions discussed herein underscore the need to embrace reticulate evolutionary patterns for understanding the timing and rates of evolutionary events, providing a clearer view of life's history.

D.4.4 The tree of life

- Spang et al. (2022) — A broad review of how genomics and phylogenomics have reshaped the tree of life, including LUCA, the root, the archaeal origin of eukaryotes and the place of viruses. *(review; CC BY 4.0)*

  Evolving Perspective on the Origin and Diversification of Cellular Life and the Virosphere. Spang et al. (2022). *Genome Biology and Evolution*. doi:10.1093/gbe/evac034

  > The tree of life (TOL) is a powerful framework to depict the evolutionary history of cellular organisms through time, from our microbial origins to the diversification of multicellular eukaryotes that shape the visible biosphere today. During the past decades, our perception of the TOL has fundamentally changed, in part, due to profound methodological advances, which allowed a more objective approach to studying organismal and viral diversity and led to the discovery of major new branches in the TOL as well as viral lineages. Phylogenetic and comparative genomics analyses of these data have, among

> others, revolutionized our understanding of the deep roots and diversity of microbial life, the origin of the eukaryotic cell, eukaryotic diversity, as well as the origin, and diversification of viruses. In this review, we provide an overview of some of the recent discoveries on the evolutionary history of cellular organisms and their viruses and discuss a variety of complementary techniques that we consider crucial for making further progress in our understanding of the TOL and its interconnection with the virosphere.

- Schleper and Rodrigues-Oliveira (2026) — A recent review of the Asgard archaea and the evidence for the two-domain tree, in which eukaryotes emerged from within the archaea. Directly relevant to the subsection on two domains and the origin of eukaryotes. *(review; CC BY 4.0)*

  Asgard archaea: have we found our microbial ancestors?. Schleper & Rodrigues-Oliveira (2026). *The EMBO Journal*. doi:10.1038/s44318-026-00719-x

  > The discovery of Asgard archaea about a decade ago has greatly reshaped our understanding of archaeal evolution and the origin of eukaryotes. Asgards are currently thought to be the closest prokaryotic relatives of eukaryotes and to represent the archaeal host lineage that participated in the endosymbiotic event leading to the first eukaryotic cell. The presence of numerous eukaryotic signature proteins in Asgard genomes supports this view and provides important insights into the deep evolutionary roots of eukaryotic cellular complexity. However, the close relationship between archaea and eukaryotes had been observed for decades, based on features that are shared in different molecular processes. This review discusses the discovery of Asgard archaea in the broader context of archaeal molecular and cellular biology and highlights how earlier findings foreshadowed their emergence. Primarily targeted at newcomers to the field, the review provides an overview of evolutionary innovations across the Archaea domain and discusses molecular and cellular features of cultivated Asgard strains in light of previous archaeal research.

- Puigbò et al. (2019) — Describes how thousands of prokaryotic gene trees can be compared to find a central tree-like signal despite horizontal gene transfer. Relevant to the subsection on whether the tree of life is a tree or a network. *(book chapter; CC BY 4.0)*

  Genome-Wide Comparative Analysis of Phylogenetic Trees: The Prokaryotic Forest of Life. Puigbò et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_8

  > Genome-wide comparison of phylogenetic trees is becoming an increasingly common approach in evolutionary genomics, and a variety of approaches for such comparison have been developed. In this article we present several methods for comparative analysis of large numbers of phylogenetic trees. To compare phylogenetic trees taking into account the bootstrap support for each internal branch, the boot-split distance (BSD) method is introduced as an extension of the previously developed split distance (SD) method for tree comparison. The BSD method implements the straightforward idea that comparison of phylogenetic trees can be made more robust by treating tree splits differentially depending on the bootstrap support. Approaches are also introduced for detecting treelike and netlike evolutionary trends in the phylogenetic Forest of Life (FOL), i.e., the entirety of the phylogenetic trees for conserved genes of prokaryotes. The principal method employed for this purpose includes mapping quartets of species onto trees to calculate the support of each quartet topology and so to quantify the tree and net contributions to the distances between species. We describe the applications methods used to analyze the FOL and the results obtained with these methods. These results support the concept of the Tree of Life (TOL) as a central evolutionary trend in the FOL as opposed to the traditional view of the TOL as a "species tree."

The following open-access paper(s) are reproduced under their original licences.

Review Article

TRANSPARENT PROCESS OPEN ACCESS CHECK FOR UPDATES

THE EMBO JOURNAL

Asgard archaea: have we found our microbial ancestors?

Christa Schleper [1]✉ & Thiago Rodrigues-Oliveira [1,2]

Abstract

The discovery of Asgard archaea about a decade ago has greatly reshaped our understanding of archaeal evolution and the origin of eukaryotes. Asgards are currently thought to be the closest prokaryotic relatives of eukaryotes and to represent the archaeal host lineage that participated in the endosymbiotic event leading to the first eukaryotic cell. The presence of numerous eukaryotic signature proteins in Asgard genomes supports this view and provides important insights into the deep evolutionary roots of eukaryotic cellular complexity. However, the close relationship between archaea and eukaryotes had been observed for decades, based on features that are shared in different molecular processes. This review discusses the discovery of Asgard archaea in the broader context of archaeal molecular and cellular biology and highlights how earlier findings foreshadowed their emergence. Primarily targeted at newcomers to the field, the review provides an overview of evolutionary innovations across the Archaea domain and discusses molecular and cellular features of cultivated Asgard strains in light of previous archaeal research.

Keywords Archaea; Eukaryogenesis; Evolution; Eukaryotes; Metagenomics
Subject Categories Evolution & Ecology; Microbiology, Virology & Host Pathogen Interaction
https://doi.org/10.1038/s44318-026-00719-x
Received 26 November 2025; Revised 28 January 2026; Accepted 3 February 2026
Published online: 5 March 2026

Introduction

The discovery of Asgard archaea, which started only around 10 years ago, has tremendously influenced our view of archaeal evolution and that of eukaryotes. Asgards are currently considered the closest known prokaryotic relatives of eukaryotes. They are thought to have contributed to the host of the endosymbiosis that occurred with a bacterium (the pre-mitochondrion) to form the first eukaryote about two billion years ago. Hundreds of genes in Asgard archaea, previously only found in eukaryotes, seem to confirm this hypothesis and shed light on the deep origins of eukaryotic features and molecular machineries. The discovery of Asgard archaea also raises novel hypotheses about the driving forces and mechanisms that gave rise to the emergence of the first eukaryotic cell.

It is a pity that Carl Woese, the discoverer of the Archaea, could not witness the Asgard detection and missed them by just a few years. Woese sensed quickly after his finding of Archaea being a separate, third domain, that these organisms would give clues about eukaryogenesis, and he wrote several articles about early evolutionary scenarios (e.g. Woese, 1987, 2004).

Although archaeal genomes are small and circular like those of bacteria, it has already become clear since the 1980s that all archaea share astonishing similarities with eukaryotes in their molecular machineries involved in transcription, replication, translation, DNA repair, and in other complexes, indicating a shared evolutionary ancestry. This has attracted many molecular biologists over the past 40 years to study archaeal information processing. In light of the findings of these first four decades of intense molecular biological research on archaea, it becomes understandable why the recent discovery of Asgard archaea has had such a tremendous impact on the research field and beyond. As was stated in many papers and reviews, the archaea research community was almost anticipating to find a lineage that was even closer to eukaryogenesis and has transferred the archaeal inventions to eukaryotes.

This review is intended to place the discovery of Asgard archaea into the bigger context of archaeal biology and to show to what extent the close evolutionary relationship with eukaryotes was recognized well before their discovery.

The review is meant primarily for newcomers to the field of archaeal molecular and cellular biology who wish to get an overview of the evolutionary innovations found throughout the whole Archaea domain. We are often asked if there is a review that could serve as a kick start into the archaea field. But there is no single one, because the domain of archaea is huge and diverse, as are its metabolisms and cell biology. This article is meant as a first orientation to the field of Archaea with an emphasis on molecular and cellular features that were known before the year 2015, but which foreshadowed the discovery of Asgards, a lineage closer to eukaryotes than all other archaea ever investigated before. The review also discusses molecular and cellular features of the currently cultivated Asgard strains in light of these earlier discoveries.

We refer the readers to other review articles for further exciting aspects of archaea, for example on metabolism and ecological distribution (Offre et al, 2013; Baker et al, 2020; Adam et al, 2017), their adaptations to extreme environments (Valentine, 2007), their

[1]Department of Functional and Evolutionary Ecology, University of Vienna, Djerassiplatz 1, Vienna 1030, Austria. [2]Present address: Department of Biology, Aarhus University, Section of Microbiology, Ny Munkegade 114, Aarhus 8000, Denmark. ✉E-mail: christa.schleper@univie.ac.at

 © The Author(s)

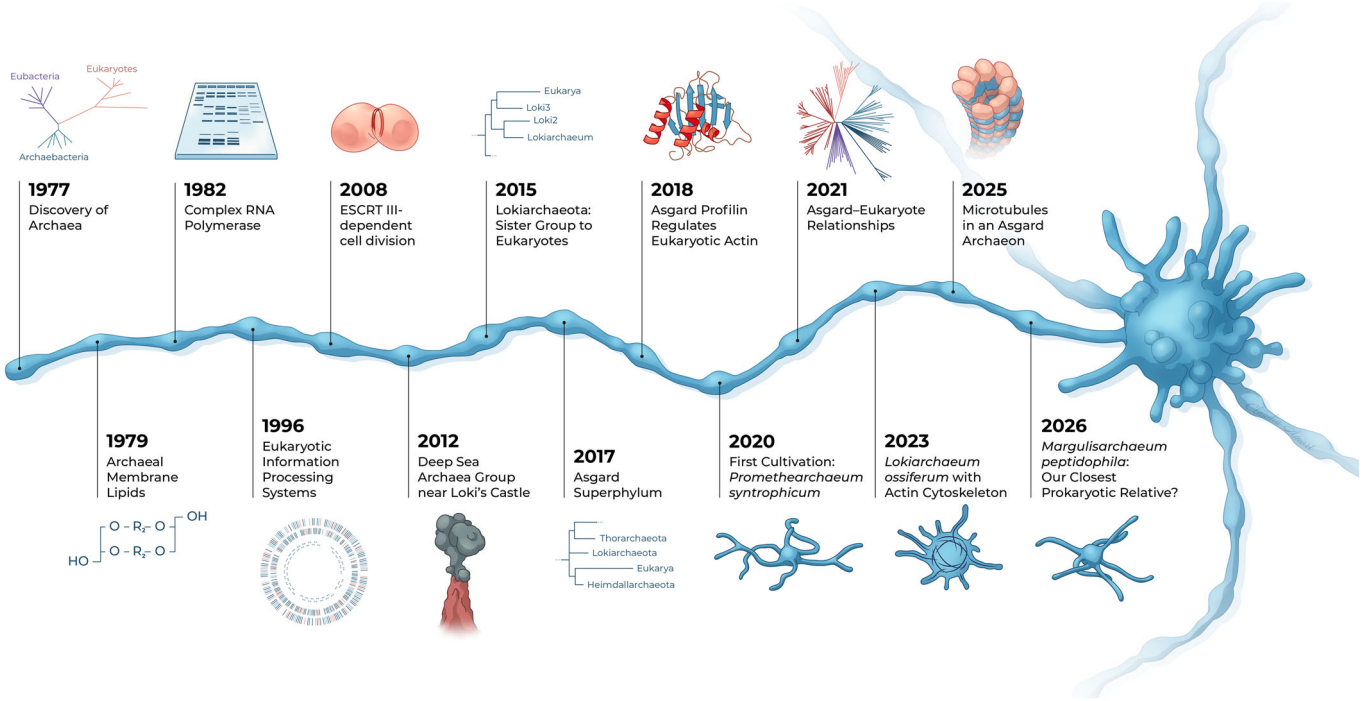


Figure 1. Timeline of some major discoveries in the archaeal field.

A few highlights from the first recognition of Archaea to the cultivation of the first Asgard representatives and their cellular structures are shown. Symbols indicate the respective methods used.

viruses (Prangishvili et al, 2017) and defense systems (Zink et al, 2020; Makarova et al, 2020), cell biology (van Wolferen et al, 2022), the comparative genomics of Asgard archaea and their eukaryotic signature proteins (Zaremba-Niedzwiedzka et al, 2017; Liu et al, 2021) as well as in-depth-discussions on different evolutionary and metabolic models for eukaryogenesis (Donoghue et al, 2023; Lopez-Garcia and Moreira, 2020), just to name a few.

Archaeal classification and taxonomy are currently subject to frequent revisions and parallel taxonomies spread (Guy and Ettema, 2011; Rinke et al, 2021; Hug et al, 2016). To facilitate orientation within the existing literature, we use the most commonly adopted terminology throughout this review, while also indicating the new GTDB nomenclature for higher ranks where needed for clarity.

From discovery to diversity

Nearly 50 years ago, Carl Woese analyzed the ribosomal RNA of prokaryotes when he unexpectedly recognized that some of the investigated organisms were as different from all other bacteria as they were from eukaryotes (Woese and Fox, 1977) (Fig. 1. He first dubbed them archaebacteria, but in 1990 introduced the concept of the three domains *Archaea*, *Bacteria* and *Eukarya* (Woese et al, 1990), because it became increasingly clear that this deep phylogenetic split was supported by many cellular features. Since the late 1970s, the unique cell membranes of *Archaea* consisting of isoprenoid-derived side chains that are linked by an ether bond to a glycerol-1-phosphate were uncovered, whereas bacteria and eukaryotes have fatty acids ester-linked to a glycerol-3-phosphate (Woese et al, 1978; Tornabene and Langworthy, 1979). In addition, archaeal lipids can form a monolayer by linking the opposing hydrophobic side chains together. The division of all living organisms into those that contain ether-linked lipids with branched side chains (archaea) and those that have ester linkages and fatty acids (bacteria and eukaryotes) has often been referred to as the lipid divide (Koga et al, 1998). This divide has puzzled researchers trying to integrate it into their various evolutionary scenarios to explain the rise of eukaryotes, but it has become less sharp recently as some archaea have the genetic capacity to form ester lipids (including some Lokiarchaea) and some bacteria ether links (Villanueva et al, 2017). It will be very interesting to reveal the nature of lipids from diverse Asgard archaea, once more biomass and cleaner cultures can be obtained.

In parallel, the group of Otto Kandler recognized early that the cell walls of archaea are quite diverse and not like those of bacteria (König and Kandler, 1979). Only a few archaeal methanogenic organisms carry a peptidoglycan cell wall, that is, however, chemically distinct from bacterial murein. Most other archaea carry a proteinaceous, almost crystalline S-layer, while some have a pseudomurein cell wall and an S-layer sheet on top (Albers and Meyer, 2011). The early finding of Otto Kandler made him a vivid proponent of Carl Woese's suggestion that archaea are indeed members of a separate, third domain (Woese et al, 1990). Interestingly, like a few other archaea (e.g., *Thermoplasma acidophilum* (Darland et al, 1970), some of the so far cultivated Asgard archaea do not seem to have a regular cell wall, but rather a more irregular coating of so far unknown composition (see below).

Among the groups of organisms that Woese originally recognized as *Archaea* were methanogens (methane producers) as

© The Author(s)

well as organisms from hot environments or hot and acidic places, i.e. (hyper)thermophilic and thermoacidophilic archaea, as well as halophiles, i.e. salt-adapted organisms (see also symbols in Fig. 2. Their different adaptations and metabolisms were subjected to many studies and despite the special growth conditions, excellent genetic models were developed for representative species of each of these three groups, including *Haloferax volcanii, Halobacterium salinarum/NERC, Methanococcus maripaludis, Methanosarcina acetivorans, Thermococcus kodakarensis, Pyrococcus furiosus, Sulfolobus acidocaldarius*, and *Saccharolobus ssp* (Leigh et al, 2011). As of today, more than 600 archaea, many (but not all) from extreme environments, have been obtained as pure cultures. Several of them define some of the physical limits of life on Earth. The record holders for the hottest temperature where cells were reported to divide, *Methanopyrus kandleri* (122 °C, (Takai et al, 2008)) and the lowest pH optimum (pH 0.7, *Picrophilus oshimae* (Schleper et al, 1995) are for example defined by archaea. Upon the discovery of Archaea as a separate group, the phylogenetic tree was quickly populated by novel lineages, as pioneers like Karl Stetter and Wolfram Zillig went out to extreme places to search for more thermophilic lineages of this novel domain (Stetter, 2013). Novel strains were also cultured from high-salt environments, while methanogenic archaea were uncovered from virtually any anoxic habitat. Biological methane formation on Earth from CO_2 or small organic acids is solely performed by methanogenic archaea, who span the widest temperature ranges from arctic temperatures to the hottest environments, but are most widespread in anoxic wetlands and sediments as well as in ruminants (Liu and Whitman, 2008) and also occur in the human gut of about half the population (Guillaume et al, 2020). Only at the beginning of this century, another ubiquitously distributed archaeal group, autotrophic ammonia-oxidizing archaea, were cultivated, which are found in virtually any aerobic habitat and play also an important role in the global nitrogen cycle and the carbon cycle (Konneke et al, 2005; Leininger et al, 2006; Brochier-Armanet et al, 2008). They are the only aerobic archaea that are found in common-place environments, including oceans, freshwater, and soils, and even on the human skin (Probst et al, 2013; Schleper and Nicol, 2010).

The phylogenetic analysis of ribosomal RNA sequences initiated by Carl Woese did not only allow the discovery of the domain *Archaea*, but it also laid the foundation for the field of microbial ecology to study the diversity of microorganisms in complex microbiomes without the need for cultivation. Until today, 16S rRNA gene sequencing is often taken as a starting point for taxonomic profiling before metagenomic techniques are used to sequence and assemble complete or almost complete genomes from environmental samples (Tyson et al, 2004). Metagenomics has led to the discovery of many new lineages and even whole new phyla within both the bacterial and archaeal domains. For example, one of the most abundant archaeal groups, the ammonia-oxidizing archaea, found in virtually all aerobic environments on Earth, was discovered by molecular techniques more than ten years before the first organisms were cultivated (DeLong, 1992; Fuhrman et al, 1992). Metagenomics also led to the discovery of the then phylum Bathyarchaeota (now TACK or Thermoproteota) (Meng et al, 2014), a widespread lineage mostly from anoxic environments of moderate temperature. The diversity of several other widespread archaeal lineages, of which we miss cultivated members until today, including e.g. many lineages of DPANN (nano-sized archaea) and Nitrososphaerales was discovered by metagenomics. Finally, in 2015 metagenomics led to the discovery of Lokiarchaea in deep marine sediments, initiating the Asgard research field (Fig. 1, Spang et al, 2015).

A complex RNA polymerase transcribes a prokaryotic genome

Already in the late 1970s Wolfram Zillig and colleagues published the first purified active RNA polymerases from halophilic and thermoacidophilic archaea, and recognized their complexity and insensitivity to some antibiotics (Zillig et al, 1979) (Fig. 2). With the isolation of more archaea, the general picture emerged that all have RNA polymerases that are as complex as those of eukaryotes with e.g. 10 out of 13 subunits being homologous (Zillig et al, 1988). Thus, archaea must have evolved this enzyme (similar to the one we carry today in our cells) quite early in evolution. These findings made Wolfram Zillig (besides Otto Kandler) the second passionate supporter of Carl Woese's proposition of Archaea being a fundamental lineage of life distinct from Bacteria and eukaryotes (Albers et al, 2013). Subsequently, several of the core transcription factors homologous to those found in eukaryotes today were described, such as TBP (TATA box binding protein), TFB (a homolog of TFIIB in eukaryotes), and often TFE (homolog of the TFIIE alpha subunit in eukaryotes), defining a minimal machinery needed for archaeal transcription initiation (reviewed in Finn Werner and Grohmann, 2011). Different from the eukaryotic counterpart, this pre-initiation complex does not require additional factors or ATP hydrolysis to form the open complex (Blombach et al, 2016). Also, no sigma factor is needed for initiation in these microorganisms in contrast to bacteria. Coherent with these findings, the core promoter elements on the DNA level are also different from those in bacteria and carry a TATA box element similar to RNA polymerase II promoters in eukaryotes (Reiter et al, 1990; Blombach et al, 2019). Transcription elongation is mediated by Spt4/5, both conserved in eukaryotes, where Spt5 is also a homolog of the bacterial factor NusG. In addition, TFS (a homolog of TFIIS in eukaryotes) helps processivity during transcription (Sanders et al, 2019). Transcription termination factor FttA (or CPSF1) is the more recently discovered element and universally present in archaea; it is an ortholog of the eukaryotic CPSF73 (Phung et al, 2013) and helps termination for genes lacking intrinsic termination sequences (Wenck and Santangelo, 2020). This demonstrates that not only the eukaryotic RNA polymerase, but also other factors associated with transcription are offsprings of early archaeal inventions (or share a common ancestry with those from archaea).

For the discussion of gene regulation in archaea, one needs to consider the prokaryotic structure of archaeal genomes. Like bacteria, archaea have small, densely packed singular and circular chromosomes with genes arranged in clusters and also often co-transcribed in operons (Allers and Mevarech, 2005). Archaeal regulatory transcription factors (aRTFs) directly bind to an operator sequence near promoter elements where they either hinder or strengthen interactions between basal TFs and promoter elements. When repressing, TFs inhibit access of the RNAP and basal TFs to the core promoter, reminiscent of bacterial systems. However, the recruitment of the transcription complex when activated might mimic eukaryotic-like mechanisms, although in eukaryotes several activator binding sites are not only located in close proximity but also at larger distances to the transcription start site (Wenck and Santangelo, 2020).

 © The Author(s)

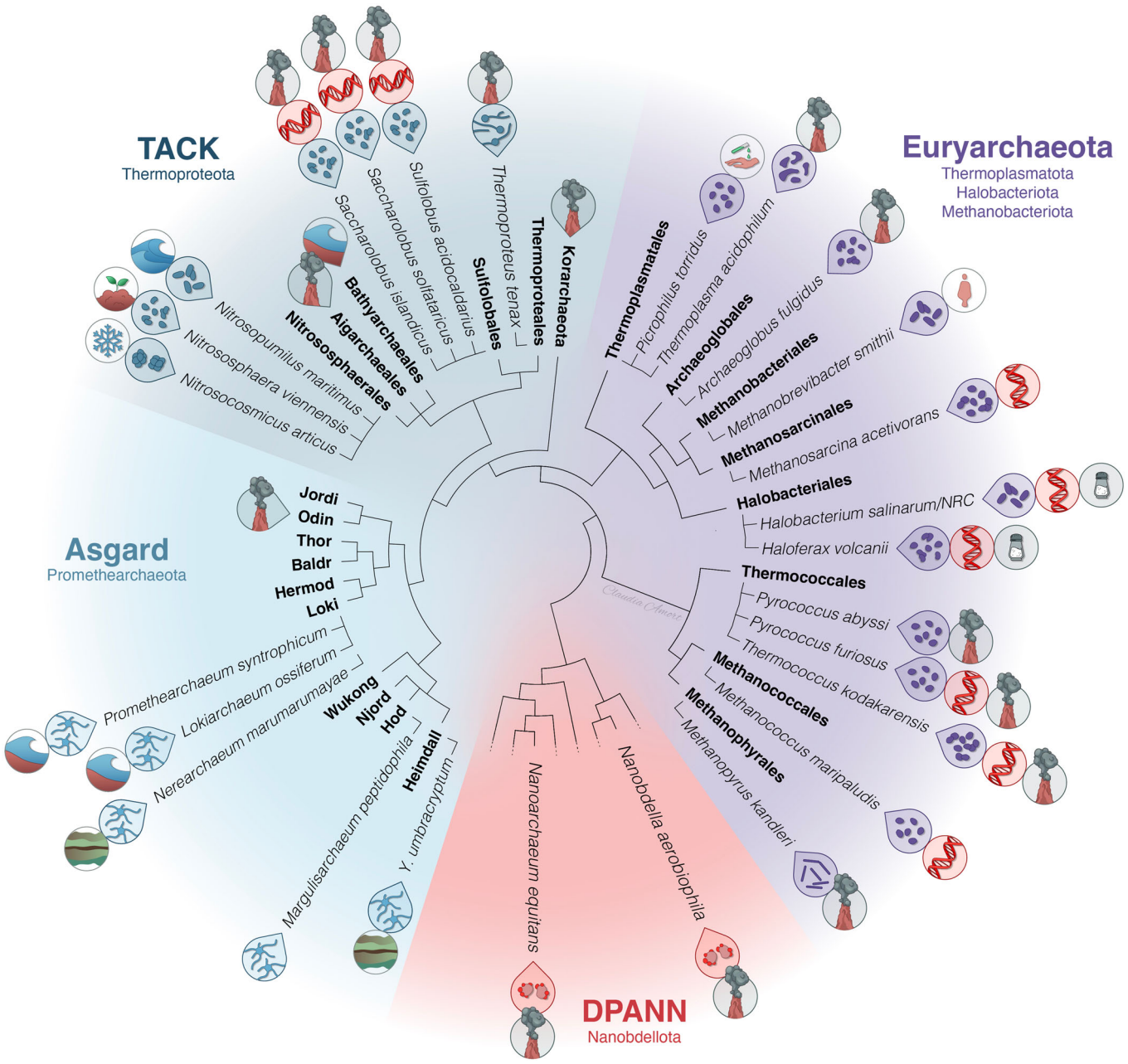


Figure 2. Phylogenetic tree of Archaea highlighting the main archaeal groups and the species referred to in the text. Only a selection of archaea and their corresponding order levels (in bold) is represented.

Archaeal species are all from laboratory cultures (not metagenomes). Outer symbols represent habitats (hot springs, acidic springs, sediments, salterns, ocean, soil, human), DNA symbol indidates genetic model organisms, inner symbols identify cell shapes. For higher level taxonomic ranks, both the most commonly used taxonomic nomenclature, like TACK, DPANN etc. (Hug et al, 2016; Zaremba-Niedzwiedzka et al, 2017), and the current GTDB nomenclature (in smaller font below) are presented. TACK Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota, DPANN Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota, Nanoarchaeota. In the case of Asgard archaea, we have avoided using rank endings (from phylum to class level) in the figure and refer to these lineages as Loki/Thor/Odin-archaea.

Diversity of archaeal chromatin

Archaea have a variety of nucleoid-associated proteins (NAPs) that organize their genomes. These likely regulate DNA compaction and thus influence gene expression. Many archaea contain structural maintenance of chromosomes (SMC) proteins that seem to act as condensins in DNA compaction and organization. Takemata et al were the first to show physical and functional compartmentalization of archaeal chromosomes, of which DNA fractions, mostly bound by coalescin, a novel archaeal-specific SMC protein, are less transcriptionally active (Takemata et al, 2019). Many archaea also contain histones that are related to those in eukaryotes (Henneman et al, 2018). The proteins are highly expressed and can cover the whole genome to form superstructures (Mattiroli et al, 2017), and their blocking effects

© The Author(s)

on transcription are overcome by transcription factors TFS and Spt4/5 (Sanders et al, 2019). Stevens et al (2020) showed that some archaeal histone paralogs might act like eukaryotic histone variants, indicating that a combinatorially complex histone-based chromatin might have evolved already in archaea.

However, in methanogens, halophiles, and ammonia-oxidizing archaea (Nitrososphaerales), histones do not have N-terminal or C-terminal extensions as known from eukaryotes, and they show little posttranslational modification (Grau-Bove et al, 2022). In contrast, Asgard archaea contain an enrichment of up to nine histone genes, some of which show extended N-terminal tails with basic amino acids (mostly lysins) that might be modifiable for regulation as known in eukaryotes. Interestingly, a recent study has indicated that certain Asgard histones assemble into "closed" and "open" hypernucleosomes, revealing a conserved archaeal chromatin conformation alongside an Asgard-specific, eukaryote-like state and thereby establishing the first structure-based model of Asgard chromatin (Ranawat et al, 2025). The full role of Asgard histones in gene regulation and chromatin organization still awaits being revealed.

The deep roots of the eukaryotic replication machinery

Shortly after the discovery of the Archaea, Forterre et al realized that some halophilic organisms are inhibited by aphidicolin, the eukaryotic inhibitor of replication (Forterre et al, 1984). The following years were then characterized by efforts to isolate archaeal DNA polymerases from hyperthermophiles to be used in the novel technology of polymerase chain reaction, PCR (Rossi et al, 1986). B family archaeal polymerases have, different from their thermostable bacterial counterparts, a 3'–5' proofreading activity and are thus less error prone and have since then been used as proofreading polymerases in PCR technology. Only later it was discovered that archaea contain another polymerase (Pol D (Ishino et al, 1998) that -at least in some species- is the only replicating protein (Cubonova et al, 2013). With the first archaeal complete genome sequences being produced in the 1990s (i.e., *Methanocaldococcus jannaschii*, back then *Methanococcus* (Bult et al, 1996), *Archaeoglobus fulgidus* (Klenk et al, 1997), *Pyrococcus horikoshii* (Kawarabayasi et al, 1998)), it became clear that the accessory replication proteins are mostly (though not exclusively) homologous to those already well-known from eukaryotes and also exhibited similar biochemical properties. Among these were the proliferating cell nuclear antigen (PCNA), replication factor C (RFC), and proteins involved in lagging-strand processing, such as DNA ligase and Fen1 (Ishino and Ishino, 2012). Similarly, the initiation of replication involves proteins that already evolved in archaea and are found in eukaryotes, in particular the minichromosome maintenance (MCM) helicase (Sakakibara et al, 2009) and the initiation protein Cdc6, also referred to as Cdc6/Orc1 (Costa et al, 2013). The archaeal helicase is a hexamer, but differently from eukaryotic cells is made up of a single protein (Barry and Bell, 2006). In *Sulfolobus*, Orc1-1 binds as a monomer to elements of the replication origin oriC1, where the ATP-bound form of the protein carries out both DNA binding and helicase recruitment functions (Samson et al, 2016). Despite differences among lineages, the archaeal replication machinery can generally be seen as a simplified, and probably ancestral form of that in eukaryotes, very similar to the transcription machinery (Barry and Bell, 2006). A recent comparative investigation of Asgard genomes revealed a structural diversification of replisomes in different lineages, which contributed to the more sophisticated machinery in eukaryotes, likely due to horizontal gene transfers on the evolutionary path from the first towards the last Eukaryotic Common Ancestor (LECA) (Feng et al, 2025). Those transferred elements included a DNA polymerase δ-like complex in baldrarchaea, a primase complex in sif/wukong/heimdallarchaea and a RFC clamp-loader complex in lokiarchaea but also RfcS and Fen1.

Considering their small circular genomes (~0.5–6 Mbp) (Kellner et al, 2018), archaea appear as an evolutionary hybrid between bacteria and eukaryotes, with a prokaryotic-type cell and a eukaryotic-type replisome. This is also true for the number of their replication origins, which range between one (as in bacteria) to several (Samson et al, 2011), as does the ploidy of genomes (between one copy to over 20) (Ludt and Soppa, 2019; Breuert et al, 2006). Interestingly, growth and replication in halophilic (and a few other) archaea are possible when all origins of replication are deleted, and growth is even accelerated under these conditions (Hawkins et al, 2013).

While their replication machinery bears a striking similarity to that in eukaryotes, archaea show an interesting and complex mosaic of distribution of DNA repair enzymes, and also encode unique ones. For example, while a canonical bacterial mismatch repair pathway based on MutL-MutS is only found in a few lineages of mesophiles, including Asgards (White and Allers, 2018), a novel enzyme, EndoMS (for endonuclease mismatch specific), was discovered in several lineages of thermophiles to serve that role (Ishino et al, 2016). Similarly, mostly mesophilic archaea (including Asgards) seem to have adopted the (bacterial) UvrABC-based nucleotide excision repair system, while most archaea, including thermophiles (both from the Euryarchaeota and TACK) as well as mesophiles including Asgards also carry the eukaryote-type NER proteins XPF, XPD, and XPB-Bax1. Similarly, a patchy distribution of enzymes is found for the base excision and double-strand break repair systems in archaea. End-joining pathways have only been identified in specific organisms, with NHEJ being found exclusively in *Methanocella paludicola* (Bartlett et al, 2013) and MMEJ being identified *in Haloferax volcanii* (Delmas et al, 2009), (Stachler et al, 2017), and *Sulfolobus islandicus* (Zhang and Whitaker, 2018). Homologous recombination (HR) is the best-studied double-strand break repair pathway in archaea (White, 2011). While the HR presynapsis proteins Mre11 and Rad50 are conserved in archaea (Liu et al, 2016; Sung et al, 2014), the archaeal HR synapsis SSBs, although present in different groups, are structurally different. In organisms classically referred to as euryarchaea, they resemble the eukaryotic RPA, whereas in the then called crenarchaea (now within TACK or Thermoproteota, Fig. 1), they are more similar to bacterial SSBs (Lin et al, 2008; White and Allers, 2018). Notably, the HR post-synapsis helicase Hel308 is found in archaea and metazoans but not in bacteria or yeast (Woodman and Bolt, 2009; Kelman and White, 2005; White and Allers, 2018).

Ribosomes and translation in archaea

Similar to the trends observed in other aspects of their molecular biology, the ribosomal architecture and translation machinery of

 © The Author(s)

archaea also share a noteworthy number of features with those of eukaryotes. By the 1990s, it was already established that, although archaeal ribosomal proteins were organized in "bacteria-like" operons, they were more closely related at the sequence level to those of eukaryotes (Olsen and Woese, 1997). Over the following decades, additional "eukaryote-like" traits were identified in archaeal ribosomes (reviewed in (Londei and Ferreira-Cerca, 2021)). One such trait involves the structure of ribosomal RNA. In contrast to canonical prokaryotic rRNAs, most eukaryotic rRNAs contain extension segments of variable sizes inserted into the universal rRNA core (Gerbi, 1996; Bowman et al, 2020). While recent studies have detected these elements in both bacteria and archaea (Penev et al, 2020; Tirumalai et al, 2020; Stepanov and Fox, 2021), it is worth noting that large segments comparable in size to those found in eukaryotes have so far been reported only in members of the Asgard archaea (Penev et al, 2020). These extensions were used to identify Asgard cells in a growing mixed culture analyzed via Cryo-electron tomography (Rodrigues-Oliveira et al, 2023). Archaeal ribosomes not only contain the universally conserved proteins found in all domains of life (Melnikov et al, 2012), but also share a significant subset exclusively with eukaryotes while lacking any proteins uniquely shared with bacteria (Marquez et al, 2011; Ban et al, 2014).

The archaeal translation apparatus has long been established to exhibit a combination of bacterial and eukaryotic/archaeal features (Bell and Jackson, 1998). As in Bacteria, transcription and translation in Archaea occur simultaneously, and archaeal species possess polycistronic mRNAs, enabling ribosomes to perform repeated initiation cycles on the same mRNA (Benelli et al, 2017). Both Shine-Dalgarno sequences and leaderless mRNAs are found in archaea, with the former more common in euryarchaeal species and the latter predominant in the so-called crenarchaeal organisms (Ma et al, 2002; Benelli et al, 2016). Notably, Shine-Dalgarno sequences are rare in the Lokiarchaea, reflecting a noticeable difference to bacterial and other archaeal translation initiation (Imachi et al, 2020; Schmitt et al, 2020). Furthermore, genomic analyses reveal that eukaryotic initiation factors display high similarity to their homologous counterparts in archaea, including aIF1, aIF1A, aIF2, and aIF5B (Kyrpides and Woese, 1998; Benelli and Londei, 2011; Schmitt et al, 2019). Another noteworthy parallel with eukaryotes is observed in the mRNA exit tunnel: in *Pyrococcus abyssi* (a euryarchaeon), the tunnel resembles the eukaryotic type but lacks certain features, instead relying on Shine-Dalgarno-assisted AUG recognition and missing the ribosomal protein eS26, which in eukaryotes stabilizes the mRNA (Coureux et al, 2020). Ribosomal protein S26 is, however, uniquely present in eukaryotes, members of TACK archaea, and also in Asgard archaea (Schmitt et al, 2020), lending further support to the hypothesis that eukaryotic ribosomes originated from this archaeal lineage.

Two different modes of cell division in archaea

Similar to other features previously discussed, the cell division apparatus in archaea displays noticeable diversity, with clear differences between Eury- and TACK archaea (van Wolferen et al, 2022). Most euryarchaea possess two homologs of the bacterial cell-division protein FtsZ (Ithurbide et al, 2022). The role of FtsZ1 and FtsZ2 in archaeal cytokinesis has been demonstrated in the model haloarchaeon *Haloferax volcanii* through GFP-based localization studies, and the deletion of either FtsZ homolog via genetic manipulation results in cytokinesis defects (Liao et al, 2021). Furthermore, FtsZ-based division mechanisms have also been investigated in *Methanobrevibacter smithii* (Pende et al, 2021). In both these organisms, it was revealed that a SepF homolog plays a role in anchoring the Z-ring to the membrane, similar to bacteria (Nussbaum et al, 2021; Pende et al, 2021).

In contrast to Euryarchaeota, another cell division mechanism, the Cdv system, is found in some members of TACK archaea and in Asgard archaea (Lindas et al, 2008; Zaremba-Niedzwiedzka et al, 2017) and is homologous to the eukaryotic ESCRT-III complex (reviewed in (Blanch Jover and Dekker, 2023)). The Cdv system was first described in *Sulfolobus acidocaldarius* (Lindas et al, 2008; Samson et al, 2008), where its core components, CdvA, CdvB, and CdvC, are encoded in a gene cluster essential for its viability (Zhang et al, 2018). CdvA is a protein unique to archaea that interacts with CdvB, a eukaryotic ESCRT-III homolog (Samson et al, 2011), whose paralogues have recently been implicated in membrane deformation and scission, mirroring the role of ESCRT complexes in eukaryotes (Tarrason Risa et al, 2020). CdvC, a homolog of eukaryotic Vps4 (Obita et al, 2007), appears to supply the energy required to depolymerize ESCRT-III filaments at the membrane, remodel their structure, and mediate membrane constriction and fission, also in analogy to the eukaryotic machinery (Schoneberg et al, 2018). Members of the *Nitrososphaeria* class (formerly Thaumarchaeota), also part of the TACK group, encode both the Cdv system and tubulin homologs related to FtsZ in their genomes (Bernander and Ettema, 2010; Yutin and Koonin, 2012). However, it has been suggested that the thaumarchaeon *Nitrosopumilus maritimus* relies on the Cdv system for cell division (Pelve et al, 2011). Its FtsZ-like proteins are distantly related to canonical FtsZ and lack the catalytically active domain required for this process, leaving their biological role unresolved (Aylett and Duggin, 2017). Asgard archaea also harbor genes for the Cdv machinery, where their CdvB proteins are even more similar to eukaryotic ESCRT-III than those of TACK archaea (Blanch Jover and Dekker, 2023). Interestingly, they also possess FtsZ homologs (Imachi et al, 2020; Ithurbide et al, 2022; Rodrigues-Oliveira et al, 2023), resulting in open questions on how each system contributes to the cell division process.

The archaeal cell cycles, some with a regulated progression

Compared with bacteria and eukaryotes, the archaeal cell cycle remains poorly characterized, with the available evidence suggesting that archaea combine features of both eukaryotes and bacteria (Cezanne et al, 2024). Early studies of *Sulfolobales* revealed an ordered cell cycle resembling that of eukaryotes, characterized by a short G1 phase and a prolonged G2 phase in which cells contain two interlinked genome copies (Bernander and Poplawski, 1997; Lundgren et al, 2008). This extended G2 phase has been proposed to facilitate DNA repair (Rhind and Russell, 1998). However, other TACK archaea, such as *Nitrosopumilus maritimus*, exhibit G1 and G2 phases of similar duration (Pelve et al, 2011). Despite evidence for regulated progression, these archaea that exhibited an ordered

© The Author(s)

cell cycle lack clear homologs of core eukaryotic cell cycle regulators, including cyclins, CDKs, and the APC/C (Cezanne et al, 2024). In contrast, archaea from the Euryarchaeota branch (halophiles and methanogens) show less defined cell cycle phases and are frequently polyploid, with genome copy numbers varying by species and growth conditions (Breuert et al, 2006; Hildenbrand et al, 2011). In haloarchaea, genome replication occurs throughout the cell cycle, suggesting the absence of a distinct S phase (Zerulla and Soppa, 2014), while methanogens often display loose cell cycle control and variable ploidy (Malandrin et al, 1999; Majernik et al, 2005). No data are yet available on the cell cycle organization of Asgard archaea, reflecting their recent discovery and limited cultivation. Understanding cell cycle regulation in this group remains an important direction for future research.

Discovery of the Asgard archaea by metagenomics

One of the groups discovered through 16S rRNA environmental surveys was the Deep Sea Archaeal Group (DSAG, formerly Deep Sea Hydrothermal Vent Crenarchaeotic Group (Takai and Horikoshi, 1999), (Takai et al, 2001), but also referred to as MBG-B (marine benthic group B (Vetriani et al, 1999)), commonly found mostly in marine sediments over the world. Especially high relative abundances of DSAG signatures appeared in certain sediment layers near the Loki's Castle hydrothermal vent system located in the Atlantic mid-ocean rift valley, representing even more than 50% of the microbial community (Jorgensen et al, 2012). The site was not directly influenced by geothermal heat, but rather by fallout from the vent system causing a distinct and sharp stratification of the sediments. Samples from a specific layer at about 2.500 mt water depth in about 2 mt deep sediment were the source for the initial discovery of the "Lokiarchaeota" (DSAG). Assembled metagenomes were sequenced for the first time, placing them as a direct sister group to eukaryotes in phylogenomic analyses (Spang et al, 2015). Equally striking was the discovery that their genomes encoded a large number of proteins previously considered unique to eukaryotes, known as Eukaryotic Signature Proteins (ESPs). These include proteins involved in membrane remodeling and vesicular trafficking, such as ESCRT system and small Ras superfamily GTPases, as well as a ubiquitinylation system and proteins linked to dynamic cytoskeleton formation, including actin, profilin, and gelsolin as well as many more functions (Spang et al, 2015). Interestingly, a few of these protein genes were detected earlier in genomes scattered throughout the archaeal tree and referred to as the "dispersed" eukaryome" (Koonin and Yutin, 2014), like e.g. the ubiquitin system in the called Aigarchaeota (Nunoura et al, 2011), as well as ESCRT-III proteins (Bernander and Ettema, 2010) and actin-related proteins (Izore et al, 2016)) in lineages of the TACK group.

Since the first discovery from the North Atlantic, additional environmental surveys have uncovered more archaeal genomes related to Lokiarchaeota. Among the newly identified groups were the Thorarchaea (Seitz et al, 2016), Odinarchaea (Zaremba-Niedzwiedzka et al, 2017), Helarchaea (Seitz et al, 2019), Heimdallarchaea (Zaremba-Niedzwiedzka et al, 2017), Gerdarchaea (Cai et al, 2020), Kariarchaea (Liu et al, 2021), Wukongarchaea (Liu et al, 2021), Hodarchaea (Liu et al, 2021; Eme et al, 2023), and others. Collectively, these organisms are now commonly referred to as the Asgard archaea, after the home of the Nordic gods (Zaremba-Niedzwiedzka et al, 2017), with their known diversity and phylogeny continuing to expand. As taxonomy developed, Lokiarchaeota was reclassified as *Lokiarchaeia* to reflect its rank at the class level (Rinke et al, 2021; Sun et al, 2021), and Promethearchaeota has been proposed for the whole phylum, after the first cultured organism *Promethearchaeum syntrophicum* ((Imachi et al, 2020), see below). The defining feature uniting all Asgard lineages is the high abundance of shared ESPs and their monophyly, with eukaryotes in most cases (but not always) emerging close to Heimdallarchaea or Hodarchaea (Liu et al, 2021; Eme et al, 2023; Da Cunha et al, 2022).

Although first discovered in deep marine sediments, Asgard archaea are now recognized as globally distributed across a wide range of ecosystems. They occur in various anoxic sediment environments (Bulzu et al, 2019; Zou et al, 2020); Hager et al, 2025; Manoharan et al, 2019) as well as in soils and rhizospheres (Cai et al, 2021), hot springs (Zaremba-Niedzwiedzka et al, 2017), hydrothermal vents (Wu et al, 2022; Rambo et al, 2022), permafrost (Liang et al, 2021), surface oceans (Rodriguez et al, 2020), and epipelagic sediments (Appler et al, 2026). Lokiarchaea, Thorarchaea are the most widely distributed groups (Manoharan et al, 2019; Cai et al, 2021; Hager et al, 2025), whereas others appear more habitat-specific, with Odinarchaea and Njordarchaea being primarily found in high-temperature environments (Zaremba-Niedzwiedzka et al, 2017; Xie et al, 2022).

This broad distribution is mirrored by considerable metabolic diversity within this group. Genomic studies have suggested not only heterotrophic potential in the Lokiarchaea (Spang et al, 2019; Zhang et al, 2021) but also varying metabolic capabilities, including lignin and humic acid degradation, CO_2 assimilation, heterotrophic lactate degradation, and aromatic compound degradation (Yin et al, 2021). Helarchaea encode methyl-CoM reductase-like enzymes, suggesting potential for hydrocarbon oxidation (Seitz et al, 2019). Thorarchaea appear to share some metabolic traits with Lokiarchaea, supporting a peptide-based heterotrophy (Seitz et al, 2016; Liu et al, 2018). They also encode ribulose bisphosphate carboxylase–like proteins (without RuBisCO activity) and nearly a complete Calvin–Benson–Bassham cycle, indicating potential use of both organic and inorganic carbon (Liu et al, 2018). Wukongarchaea appear to consist of obligate hydrogenotrophic acetogens with a chemolithotrophic lifestyle (Liu et al, 2021). Heimdallarchaea possess a versatile metabolic repertoire, with evidence for a heterotrophic lifestyle via fermentation, anaerobic respiration, or even aerobic respiration (Spang et al, 2019; Appler et al, 2026). Their genomes encode a complete tricarboxylic acid (TCA) cycle supported by an electron transport chain containing V/A-type ATPase, succinate dehydrogenase, NADH–quinone oxidoreductase, and cytochrome c oxidase (Bulzu et al, 2019).

While most of the initial insights into Asgard biology stemmed from genomic data, more recent studies have begun to investigate properties of Asgard ESP proteins through biochemical experiments. Profilins from Lokiarchaea and Odinarchaea were demonstrated to adopt the typical profilin fold and interact with rabbit actin, as well as being capable to induce polymerization of mammalian actin (Akil and Robinson, 2018). Profilins from Heimdallarchaea were also shown to inhibit actin polymerization, highlighting a potential regulatory role (Survery et al, 2021), as did Thorarchaeal profilins (Inturi et al, 2022) and gelsolins

 © The Author(s)

(Akil et al, 2020). ESCRT-III and VPS4 were shown to possess chromatin-binding properties (Nachmias et al, 2023), and ESCRT-III proteins from Lokiarchaea self-assembled into helical filaments, hallmarks of the ESCRT system, that bound and deformed eukaryotic-like membrane vesicles (Melnikov et al, 2025; Souza et al, 2025). Furthermore, it has also been demonstrated that Asgard ESCRT-IIIB and ESCRT-IIIA form functionally partitioned polymers whose sequential assembly and structural transitions recapitulate the conserved membrane-deformation pathway that later evolved in eukaryotes (Souza et al, 2025). Together, these findings were extensions from genome-based predictions to experimental evidence providing the first functional insights into Asgards prior to their isolation.

Asgard archaea in laboratory culture

In 2020, Imachi and Nobu, together with colleagues from the Jamstec Institute in Japan, published the first culture of an Asgard archaeon, which they named *Candidatus* Promethearchaeum syntrophicum (Imachi et al, 2020). It is an anaerobic heterotrophic member of the Lokiarchaea which was maintained in co-culture with hydrogen-consuming microorganisms and others; by now it has been obtained as a clean co-culture with only one archaeal methanogenic partner (Imachi et al, 2024). Isolated from deep marine sediments in Japan's Kumano region, *P. syntrophicum* required a long and labor-intensive cultivation effort, mostly because of its slow growth, a common feature of deep-sea microbes, which includes long lag phases and a doubling time of 14–25 days. The morphology of *P. syntrophicum* was most striking and never observed before in archaea or bacteria: a small spherical central cell body with long—sometimes branching—protrusions (Fig. 3A–C). While many researchers had expected some internal complexity in Asgard cells based on their genomic content, the complexity of these cells seemed rather to be manifested in their cell shape.

With the cultivation of a second strain, *Candidatus* Lokiarchaeum ossiferum, which was also enriched with hydrogen-consuming microorganisms but from a shallow urban estuarine canal in Piran, Slovenia (Rodrigues-Oliveira et al, 2023) the first internal structures came to light. Although still slow-growing, *Ca*. L. ossiferum displayed a faster growth rate than *P. syntrophicum*, with a doubling time of 7–14 days and achieving up to 100-fold higher cell densities, which allowed for detailed structural investigations. Like *P. syntrophicum*, *Ca*. L. ossiferum possesses a spherical cell body with protrusions; however, these protrusions appeared more irregular, frequently branched, and often constricted, displaying bulbous structures along their length or at the tips (Fig. 3D–F). Cells were imaged with Cryo-electron tomography, identifying *Ca*. L. ossiferum cells based on characteristic expansion segments of their ribosomes. Cells were surrounded by a single membrane and complex surface structures, but no S-layer (Fig. 3F). A long-range cytoskeleton was found in the cell bodies, protrusions, and constrictions with twisted double-stranded filaments consistent with F-actin. It was confirmed to consist of the most highly expressed actin, more precisely lokiactin, highly conserved in all Asgard genomes. In a follow-up study, additional cytoskeletal elements, namely microtubules, were described (Wollweber et al, 2025). Heterologous expressions and cryo-electron tomography structures demonstrated that archaeal tubA/B form eukaryote-like heterodimers, which assembled into 5-protofilament bona fide microtubules in vitro. Non-canonical microtubules with 7-protofilaments formed through heterodimers with an additional paralogue (Wollweber et al, 2025). Tubulins are so far only found in few Asgard genomes, thus not representing a general feature of the group. However, the study suggested a pre-eukaryotic origin of microtubules as for the actin cytoskeleton (Wollweber et al, 2025).

It is noteworthy that the first cultivated Lokiarchaea belong to the same clade, while the most environmentally widespread lineages still remain uncultured (Yin et al, 2021; Hager et al, 2025). Nevertheless, a few novel isolates have recently been obtained in culture, such as *Ca*. Margulisarchaeum peptidophila, the first hodarchaeal member, which is considered one of the closest lineages to eukaryotes (Imachi et al, 2025). In addition, a novel member of the Loki-branch (Nobs et al, 2025) and one of the Heimdall-branch (MacLeod et al, 2025) have been reported (see Fig. 2). While it may be tempting to assume from these cultures that all Asgard archaea share the same morphology, targeted fluorescence in situ hybridization experiments with environmental samples have suggested a variety of cell shapes, indicating that the morphological diversity is perhaps greater than what current cultures might suggest, or that cells adopt different shapes according to their growth state or environment (Fig. 3G–I) (Avci et al, 2022; Avci et al, 2025).

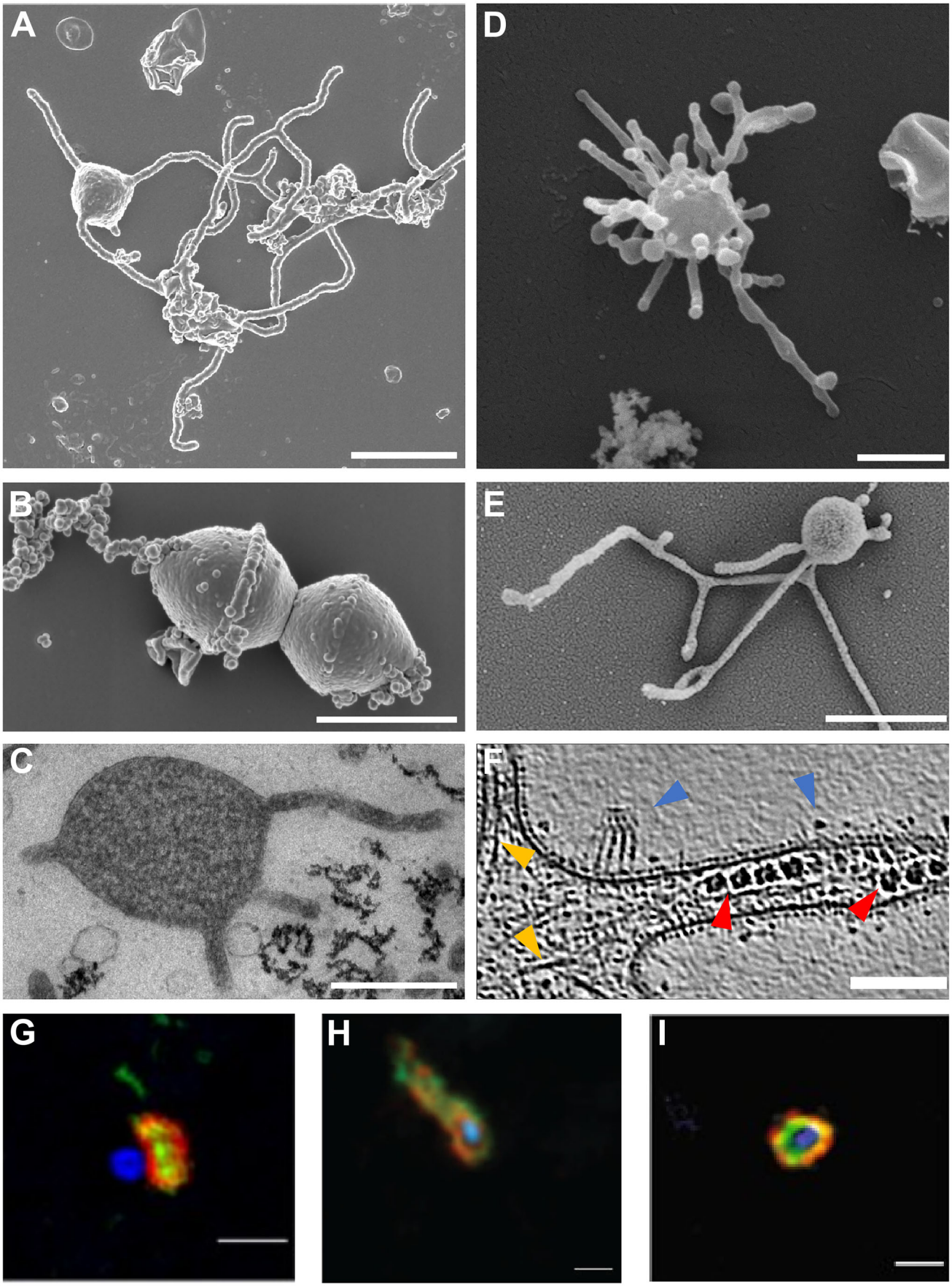
The first cultivated organisms demonstrate that fundamental features such as cell morphology, the organization of cytoskeletal elements, and many other cellular processes like cytokinesis cannot be inferred from sequence data only. Some of these features now seen in cultivated isolates were potentially crucial in eukaryogenesis. In particular, the cytoskeleton could have played an essential role in the merging of cells by phagocytosis (or another mechanism). With intracellular compartments, cells also became bigger, and a regulated intracellular transport that is also based on a cytoskeleton must have increasingly replaced mere diffusion. Equally, cell division of larger, complex cells requires a dynamic cytoskeleton.

Models for eukaryogenesis

The origin of Eukaryotes has long been debated, and many evolutionary models have been developed reflecting ongoing scientific progress. While complex multicellularity is believed to have emerged independently several times, eukaryogenesis itself is thought to have occurred only once in evolutionary history, as eukaryotes form a monophyletic group that shares a single last eukaryotic common ancestor (LECA) (Williamson et al, 2025). However, eukaryogenesis might date back well before LECA, and might have happened more often when the first events occurred forming FECA (the first eukaryotic common ancestor) that led to a suit of follow-up events.

The idea that mitochondria and chloroplasts originated from endosymbiosis dates back to the early 20th century (Mereschkowsky, 1910; Wallin, 1883, 1922; Ward, 1883; Martin and Kowallik, 1999), was revived and popularized by Lynn Margulis in the 1960s (Sagan, 1967), and then formally proved by small subunit RNA analyses by Woese (Woese and Fox, 1977). When Woese proposed the concept of the three domains of life (Woese et al, 1990) he imagined an extra lineage that led to a proto-eukaryote (a relatively complex cell with a nucleus but no mitochondria yet) which formed a sister lineage to archaea and would later engulf a bacterial endosymbiont. Opposed to that,

© The Author(s)



 © The Author(s)

Figure 3. Morphological features of Asgard archaea cells.

(**A**, **B**) Scanning electron micrographs of *Promethearchaeum syntrophicum* (Imachi et al, 2020) with potentially dividing cells in (**B**). (**C**) Transmission electron micrograph of an ultrathin section of *P. syntrophicum* (Imachi et al, 2020). (**D, E**) Scanning electron micrographs of *Ca.* Lokiarchaeum ossiferum (personal collection). (**F**) Cryo-electron tomogram of *Ca.* L. ossiferum. Blue arrows indicate the unusual cell surface structures of this organism, orange arrows highlight actin filaments, and red arrows indicate ribosomes (adapted from (Rodrigues-Oliveira et al, 2023). (**G–I**) CARD-FISH analyses from cells collected in Aarhus Bay, Denmark of (**G**) Lokiarchaeal cells, Red—probe Lok1183, Green— Probe Lok1378, Blue—DAPI (Avci et al, 2022), (**H**) Heimdallarchaeal cells, Red—probe Heim529, Green—Probe Heim329, Blue—DAPI (Avci et al, 2022), and (**I**) Hodarchaeal cells, Red—probe Hod193, Green—Probe Hod286, Blue—DAPI (Avci et al, 2025). Scale bars: 1 µm (**A,B,D,E,G,H,I**), 500 nm (**C**), and 100 nm (**F**).

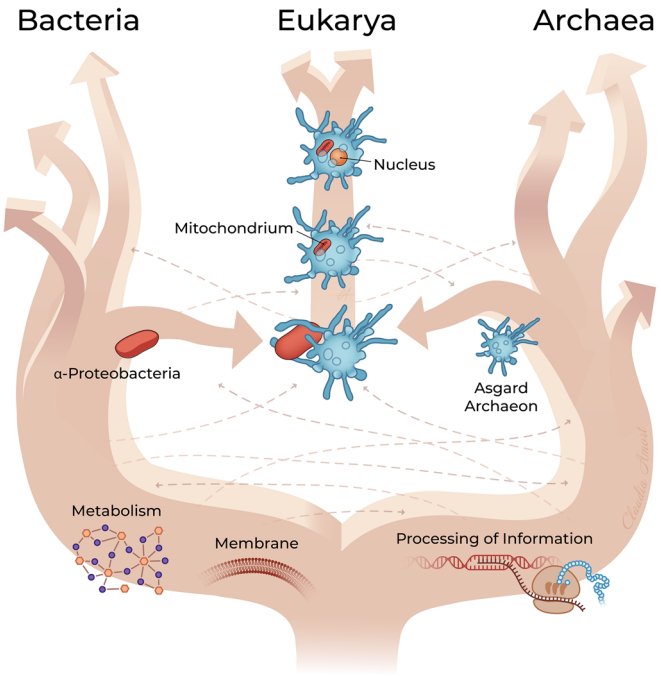


Figure 4. Visualization of a current model about the origin of eukaryotes.

The central role of Asgard archaea in this process is highlighted and the early split of Bacteria and Archaea as outlined in this review. Thin lines represent putative horizontal gene transfers between domains. The reader is encouraged to consult recent reviews and papers to appreciate the full breadth of ideas and hypotheses on the process of eukaryogenesis.

different models were already proposed in the early 1980s when ribosomal features and RNA polymerases revealed the close archaea–eukaryote link. They suggested archaea played a direct and central role in eukaryogenesis (Henderson et al, 1984; Lake et al, 1984; Zillig et al, 1985). When phylogenomic analyses proved that amitochondriate eukaryotes derive from mitochondriate ones and are thus not from an ancestral (proto-eukaryotic) lineage that could have supplied the host part in endosymbiosis, the model in which modern eukaryotes evolved from the merger of two prokaryotic cells became again more popular. The idea that eukaryogenesis coincided with endosymbiosis of an archaeon engulfing a bacterium that later evolved into mitochondria is now the most popular (Fig. 4) (Embley and Martin, 2006; Cox et al, 2008; Guy and Ettema, 2011; Williams et al, 2012; Koonin and Yutin, 2014). Several models have attempted to explain the nature of this symbiotic interaction and the driving forces for endosymbiosis (Lopez-Garcia and Moreira, 2023). One example is the Hydrogen hypothesis, which proposes that the initial symbiosis was driven by the transfer of H_2 from the symbiont to the host (Martin and Muller, 1998). Another hypothesis, the Ox-Tox model, suggests instead that the symbiont consumed oxygen toxic to the host (Kurland and Andersson, 2000). These models continue to be revised as new data emerge, and in some cases also involve two bacteria (a deltaproteobacterium and an alphaproteobacterium) and an archaeal partner that produces hydrogen (Syntrophy hypothesis) (Moreira and Lopez-Garcia, 1998; Lopez-Garcia and Moreira, 2020). Currently, it is widely believed that the identity of the endosymbiont that led to mitochondria stems from alphaproteobacteria or a closely related lineage (Yang et al, 1985; Carvalho et al, 2015; Garcia et al, 2023). Although sometimes debated (Zhang et al, 2025; Da Cunha et al, 2022), phylogenomics analyses and the conservation of ESPs in Asgard archaea now overwhelmingly support these as the archaeal partner. The exact placement of eukaryotes within the Asgard lineage is still a moving target, given the regularly changing view of their diversity and phylogeny (Zhang et al, 2025; Liu et al, 2021; Eme et al, 2023). The exact placement of Eukaryotes may also not be necessarily relevant for understanding eukaryogenesis: Whereas a whole pool of evolutionary inventions which happened within the Asgard group might have contributed to the emergence of the LECA about 2 billion years ago, they might not all be found in today's living closest sister group to eukaryotes (see for example (Wu et al, 2022)). Novel evolutionary models take into account that the complexity of the cells with their protrusions might have mediated the close interaction and eventually engulfment of the symbiotic partner (Baum and Baum, 2014; Tobiasson et al, 2026; Imachi et al, 2020). The observed flexibility of the long protrusions of Asgard cells (Imachi et al, 2020; Rodrigues-Oliveira et al, 2023), but also recently observed cellular dynamics and ameboid-like cell migration, seems to support such models (Radler et al, 2025).

The future ahead

The discovery of Asgard archaea involved a suite of observations and persistent research that started with the recognition of novel deep-branching archaea by 16S rRNA signatures. But why was their discovery as the potential eukaryotic ancestor missed for so long? In most environments preferred by Asgard archaea, they occupy no more than a few % of the microbiome diversity. So it seems it was luck that such a naturally high enrichment of more than 50% Asgard (then termed DSAG) sequences was discovered in certain layers of an unusually stratified sediment near the hydrothermal vent system Loki's castle (Jorgensen et al, 2012), as this in turn allowed the successful assembly of the very first metagenomes (Spang et al, 2015). Then the discovery of these Loki signatures in their enrichment reactors encouraged researchers to focus on the

© The Author(s)

cultivation of this particular lineage. In total, it took only ten years from the first identification of Lokiarchaeal genomes to the assembly of hundreds of Asgard genomes in metagenomic studies and the first cultivated organisms as well as in vitro studies of ESP proteins.

This research field has become vibrant with many activities now in different fields of evolution, including phylogenomics, comparative (meta)genomics, cultivation of syntrophic consortia and in the investigation of deep origins of eukaryotic features within Asgards. The near future will certainly reveal many more details about this still enigmatic group that is worth studying not only in the light of eukaryotic evolution but also in itself. It is already becoming possible to grow cultures to high enough biomass for biochemical investigations, and we will hopefully be able to develop genetic systems for these organisms to allow studying gene functions in vivo. Through the cultivation of more species, we will better understand which -sometimes unexpected- features of Asgards are common to the whole phylum and are thus likely to be transferred to the first eukaryotic cell and which ones are unique to certain lineages.

Could eukaryogenesis still happen today? Nobody has currently the answer to that, but Charles Darwin answered the question whether life could have originated more than once: "It is often said that all the conditions for the first production of a living organism are now present, which could ever have been present. But if (and oh what a big if) we could conceive in some warm little pond with all sorts of ammonia and phosphoric salts,—light, heat, electricity, etc. present, that a protein compound was chemically formed, ready to undergo still more complex changes, at the present day such matter would be instantly devoured, or absorbed, which would not have been the case before living creatures were formed" (Darwin, 1871).

Considering how complex and delicate the first close interaction of bacteria with an Asgard archaeon and their further evolution into LECA must have been, this view might also apply to eukaryogenesis, should it occur nowadays. Then, how likely can it be that a similar evolution would occur on other planets? These deep evolutionary questions are not new, and reach out to other disciplines, such as astrobiology and philosophy. What has changed now is the fact that we are a step closer to completing the picture, and able to tell a more comprehensive evolutionary story. With Asgard archaea as the plausible second partner in the symbiosis that led to eukaryotes, we have a narrative that can be communicated to the public better than ever before. We can tell how and why this event was so unique and fascinating, and of course, crucial for our very own existence. And not to forget, as Lynn Margulis liked to point out: "it was an event of cooperation and not competition" (Margulis and Sagan, 1986).

Peer review information

A peer review file is available at https://doi.org/10.1038/s44318-026-00719-x

References

Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S (2017) The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J 11:2407–2425

Akil C, Robinson RC (2018) Genomes of Asgard archaea encode profilins that regulate actin. Nature 562:439–443

Akil C, Tran LT, Orhant-Prioux M, Baskaran Y, Manser E, Blanchoin L, Robinson RC (2020) Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. Proc Natl Acad Sci USA 117:19904–19913

Albers SV, Forterre P, Prangishvili D, Schleper C (2013) The legacy of Carl Woese and Wolfram Zillig: from phylogeny to landmark discoveries. Nat Rev Microbiol 11:713–719

Albers SV, Meyer BH (2011) The archaeal cell envelope. Nat Rev Microbiol 9:414–426

Allers T, Mevarech M (2005) Archaeal genetics—the third way. Nat Rev Genet 6:58–73

Appler KE, Lingford JP, Gong X, Panagiotou K, Leao P, Langwig MV, Greening C, Ettema TJ, De Anda V, Baker BJ (2026) Oxygen metabolism in descendants of the archaeal-eukaryotic ancestor. Nature https://doi.org/10.1038/s41586-026-10128-z

Avci B, Brandt J, Nachmias D, Elia N, Albertsen M, Ettema TJG, Schramm A, Kjeldsen KU (2022) Spatial separation of ribosomes and DNA in Asgard archaeal cells. ISME J 16:606–610

Avci B, Panagiotou K, Albertsen M, Ettema TJG, Schramm A, Kjeldsen KU (2025) Peculiar morphology of Asgard archaeal cells close to the prokaryote-eukaryote boundary. mBio 16:e0032725

Aylett CHS, Duggin IG (2017) The tubulin superfamily in Archaea. In: Löwe J, Amos LA (eds) Prokaryotic cytoskeletons subcellular biochemistry. Springer, Cham, pp 393–417

Baker BJ, De Anda V, Seitz KW, Dombrowski N, Santoro AE, Lloyd KG (2020) Diversity, ecology and evolution of Archaea. Nat Microbiol 5:887–900

Ban N, Beckmann R, Cate JH, Dinman JD, Dragon F, Ellis SR, Lafontaine DL, Lindahl L, Liljas A, Lipton JM et al (2014) A new system for naming ribosomal proteins. Curr Opin Struct Biol 24:165–169

Barry ER, Bell SD (2006) DNA replication in the archaea. Microbiol Mol Biol Rev 70:876–887

Bartlett EJ, Brissett NC, Doherty AJ (2013) Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. Proc Natl Acad Sci USA 110:E1984–E1991

Baum DA, Baum B (2014) An inside-out origin for the eukaryotic cell. BMC Biol 12:76

Bell SD, Jackson SP (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends Microbiol 6:222–228

Benelli D, La Teana A, Londei P (2017) Translation regulation: the Archaea-Eukaryal connection. In: Clouet-d'Orval B (ed) RNA metabolism and gene expression in archaea nucleic acids and molecular biology, Springer, Cham, 71–88

Benelli D, La Teana A, Londei P (2016) Evolution of translational initiation: from Archaea to Eukarya. In: Hernández G, Jagus R (eds) Evolution of the protein synthesis machinery and its regulation. Springer, Cham, pp 61–79

Benelli D, Londei P (2011) Translation initiation in Archaea: conserved and domain-specific features. Biochem Soc Trans 39:89–93

Bernander R, Ettema TJ (2010) FtsZ-less cell division in archaea and bacteria. Curr Opin Microbiol 13:747–752

Bernander R, Poplawski A (1997) Cell cycle characteristics of thermophilic archaea. J Bacteriol 179:4963–4969

Blanch Jover A, Dekker C (2023) The archaeal Cdv cell division system. Trends Microbiol 31:601–615

Blombach F, Matelska D, Fouqueau T, Cackett G, Werner F (2019) Key concepts and challenges in archaeal transcription. J Mol Biol 431:4184–4201

Blombach F, Smollett KL, Grohmann D, Werner F (2016) Molecular mechanisms of transcription initiation-structure, function, and evolution of TFE/TFIIE-like factors and open complex formation. J Mol Biol 428:2592–2606

 © The Author(s)

Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD (2020) Root of the tree: the significance, evolution, and origins of the ribosome. Chem Rev 120:4848–4878

Breuert S, Allers T, Spohn G, Soppa J (2006) Regulated polyploidy in halophilic archaea. PLoS ONE 1:e92

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6:245–252

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD et al (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073

Bulzu PA, Andrei AS, Salcher MM, Mehrshad M, Inoue K, Kandori H, Beja O, Ghai R, Banciu HL (2019) Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. Nat Microbiol 4:1129–1137

Cai M, Liu Y, Yin X, Zhou Z, Friedrich MW, Richter-Heitmann T, Nimzyk R, Kulkarni A, Wang X, Li W et al (2020) Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. Sci China Life Sci 63:886–897

Cai M, Richter-Heitmann T, Yin X, Huang WC, Yang Y, Zhang C, Duan C, Pan J, Liu Y, Liu Y et al (2021) Ecological features and global distribution of Asgard archaea. Sci Total Environ 758:143581

Carvalho DS, Andrade RF, Pinho ST, Goes-Neto A, Lobao TC, Bomfim GC, El-Hani CN (2015) What are the evolutionary origins of mitochondria? A complex network approach. PLoS ONE 10:e0134988

Cezanne A, Foo S, Kuo YW, Baum B (2024) The archaeal cell cycle. Annu Rev Cell Dev Biol 40:1–23

Costa A, Hood IV, Berger JM (2013) Mechanisms for initiating cellular DNA replication. Annu Rev Biochem 82:25–54

Coureux PD, Lazennec-Schurdevin C, Bourcier S, Mechulam Y, Schmitt E (2020) Cryo-EM study of an archaeal 30S initiation complex gives insights into evolution of translation initiation. Commun Biol 3:58

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaebacterial origin of eukaryotes. Proc Natl Acad Sci USA 105:20356–20361

Cubonova L, Richardson T, Burkhart BW, Kelman Z, Connolly BA, Reeve JN, Santangelo TJ (2013) Archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in *Thermococcus kodakarensis*. J Bacteriol 195:2322–2328

Da Cunha V, Gaia M, Forterre P (2022) The expanding Asgard archaea and their elusive relationships with Eukarya. mLife 1:3–12

Darland G, Brock TD, Samsonoff W, Conti SF (1970) A thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. Science 170:1416–1418

Darwin C (1871) Letter to J.D. Hooker. In: Hooker JD (ed) University of Cambridge, Darwin Correspondence Project. https://cudl.lib.cam.ac.uk/collections/darwinhooker/1

Delmas S, Shunburne L, Ngo HP, Allers T (2009) Mre11-Rad50 promotes rapid repair of DNA damage in the polyploid archaeon Haloferax volcanii by restraining homologous recombination. PLoS Genet 5:e1000552

DeLong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci USA 89:5685–5689

Donoghue PCJ, Kay C, Spang A, Szollosi G, Nenarokova A, Moody ERR, Pisani D, Williams TA (2023) Defining eukaryotes to dissect eukaryogenesis. Curr Biol 33:R919–R929

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–630

Eme L, Tamarit D, Caceres EF, Stairs CW, De Anda V, Schon ME, Seitz KW, Dombrowski N, Lewis WH, Homa F et al (2023) Inference and reconstruction of the Heimdallarchaeial ancestry of eukaryotes. Nature 618:992–999

Feng Y, Ding J, Lin Y, Cui D, Li K, Zheng D, Cai Z, Bell SD, Wu F (2025) Serial innovations by Asgard archaea shaped the DNA replication machinery of the early eukaryotic ancestor. Nat Ecol Evol 9:2333–2345

Finn Werner F, Grohmann D (2011) Evolution of multisubunit RNA polymerases in the three domains of life. Nat Rev Microbiol 9:85–98

Forterre P, Elie C, Kohiyama M (1984) Aphidicolin inhibits growth and DNA synthesis in halophilic arachaebacteria. J Bacteriol 159:800–802

Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. Nature 356:148–149

Garcia PS, Barras F, Gribaldo S (2023) Components of iron-Sulfur cluster assembly machineries are robust phylogenetic markers to trace the origin of mitochondria and plastids. PLoS Biol 21:e3002374

Gerbi S (1996) Expansion segments: regions of variable size that interrupt the universal core secondary structure of ribosomal RNA. In: Zimmermann RA, Dahlberg AE (eds) Ribosomal RNA: structure, evolution, processing, and function in protein biosynthesis. Telford-CRC Press, Boca Raton, pp 71–87

Grau-Bove X, Navarrete C, Chiva C, Pribasnig T, Anto M, Torruella G, Galindo LJ, Lang BF, Moreira D, Lopez-Garcia P et al (2022) A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. Nat Ecol Evol 6:1007–1023

Guillaume B, Brugère JF, Gribaldo S, Schmitz RA, Moissl-Eichinger C (2020) The host-associated archaeome. Nat Rev Microbiol 18:622–636

Guy L, Ettema TJ (2011) The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol 19:580–587

Hager K, Luo ZH, Montserrat-Diez M, Ponce-Toledo RI, Baur P, Dahlke S, Andrei AS, Bulzu PA, Ghai R, Urich T et al (2025) Diversity and environmental distribution of Asgard archaea in shallow saline sediments. Front Microbiol 16:1549128

Hawkins M, Malla S, Blythe MJ, Nieduszynski CA, Allers T (2013) Accelerated growth in the absence of DNA replication origins. Nature 503:544–547

Henderson E, Oakes M, Clark MW, Lake JA, Matheson AT, Zillig W (1984) A new ribosome structure. Science 225:510–512

Henneman B, van Emmerik C, van Ingen H, Dame RT (2018) Structure and function of archaeal histones. PLoS Genet 14:e1007582

Hildenbrand C, Stock T, Lange C, Rother M, Soppa J (2011) Genome copy numbers and gene conversion in methanogenic archaea. J Bacteriol 193:734–743

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K et al (2016) A new view of the tree of life. Nat Microbiol 1:16048

Imachi H, Nobu MK, Ishii S, Hirakata Y, Ikuta T, Isaji Y, Miyata M, Miyazaki M, Morono Y, Murata K et al (2025) Eukaryotes' closest relatives are internally simple syntrophic archaea. Preprint at https://www.biorxiv.org/content/10.1101/2025.02.26.640444v1.full

Imachi H, Nobu MK, Kato S, Takaki Y, Miyazaki M, Miyata M, Ogawara M, Saito Y, Sakai S, Tahara YO et al (2024) Promethearchaeum syntrophicum gen. nov., sp. nov., an anaerobic, obligately syntrophic archaeon, the first isolate of the lineage 'Asgard' archaea, and proposal of the new archaeal phylum Promethearchaeota phyl. nov. and kingdom Promethearchaeati regn. nov. Int J Syst Evol Microbiol 74:006435

Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, Takano Y, Uematsu K, Ikuta T, Ito M et al (2020) Isolation of an archaeon at the prokaryote-eukaryote interface. Nature 577:519–525

Inturi R, Lara S, Derweesh M, Chi CN (2022) Structural characterization of a thorarchaeota profilin indicates eukaryotic-like features but with an extended N-terminus. Adv Biol 6:e2101323

Ishino S, Nishi Y, Oda S, Uemori T, Sagara T, Takatsu N, Yamagami T, Shirai T, Ishino Y (2016) Identification of a mismatch-specific endonuclease in hyperthermophilic Archaea. Nucleic Acids Res 44:2977–2986

© The Author(s)

Ishino Y, Ishino S (2012) Rapid progress of DNA replication studies in Archaea, the third domain of life. Sci China Life Sci 55:386–403

Ishino Y, Komori K, Cann IK, Koga Y (1998) A novel DNA polymerase family found in Archaea. J Bacteriol 180:2232–2236

Ithurbide S, Gribaldo S, Albers SV, Pende N (2022) Spotlight on FtsZ-based cell division in Archaea. Trends Microbiol 30:665–678

Izore T, Kureisaite-Ciziene D, McLaughlin SH, Lowe J (2016) Crenactin forms actin-like double helical filaments regulated by arcadin-2. eLife 5:e21600

Jorgensen SL, Hannisdal B, Lanzen A, Baumberger T, Flesland K, Fonseca R, Ovreas L, Steen IH, Thorseth IH, Pedersen RB et al (2012) Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. Proc Natl Acad Sci USA 109:E2846–E2855

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A et al (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 5:55–76

Kellner S, Spang A, Offre P, Szollosi GJ, Petitjean C, Williams TA (2018) Genome size evolution in the Archaea. Emerg Top Life Sci 2:595–605

Kelman Z, White MF (2005) Archaeal DNA replication and repair. Curr Opin Microbiol 8:669–676

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD et al (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370

Koga Y, Kyuragi T, Nishihara M, Sone N (1998) Did archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent. J Mol Evol 46:54–63

König H, Kandler O (1979) N-Acetyltalosaminuronic acid a constituent of the pseudomurein of the genus Methanobacterium. Arch Microbiol 123:295–299

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437:543–546

Koonin EV, Yutin N (2014) The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. Cold Spring Harb Perspect Biol 6:a016188

Kurland CG, Andersson SG (2000) Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev 64:786–820

Kyrpides NC, Woese CR (1998) Universally conserved translation initiation factors. Proc Natl Acad Sci USA 95:224–228

Lake JA, Henderson E, Oakes M, Clark MW (1984) Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci USA 81:3786–3790

Leigh JA, Albers SV, Atomi H, Allers T (2011) Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. FEMS Microbiol Rev 35:577–608

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature 442(7104):806–809

Liang R, Li Z, Lau Vetter MCY, Vishnivetskaya TA, Zanina OG, Lloyd KG, Pfiffner SM, Rivkina EM, Wang W, Wiggins J et al (2021) Genomic reconstruction of fossil and living microorganisms in ancient Siberian permafrost. Microbiome 9:110

Liao Y, Ithurbide S, Evenhuis C, Lowe J, Duggin IG (2021) Cell division in the archaeon Haloferax volcanii relies on two FtsZ proteins with distinct functions in division ring assembly and constriction. Nat Microbiol 6:594–605

Lin Y, Lin LJ, Sriratana P, Coleman K, Ha T, Spies M, Cann IK (2008) Engineering of functional replication protein a homologs based on insights into the evolution of oligonucleotide/oligosaccharide-binding folds. J Bacteriol 190:5766–5780

Lindas AC, Karlsson EA, Lindgren MT, Ettema TJ, Bernander R (2008) A unique cell division machinery in the Archaea. Proc Natl Acad Sci USA 105:18942–18946

Liu Y, Makarova KS, Huang WC, Wolf YI, Nikolskaya AN, Zhang X, Cai M, Zhang CJ, Xu W, Luo Z et al (2021) Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature 593:553–557

Liu Y, Sung S, Kim Y, Li F, Gwon G, Jo A, Kim AK, Kim T, Song OK, Lee SE et al (2016) ATP-dependent DNA binding, unwinding, and resection by the Mre11/Rad50 complex. EMBO J 35:743–758

Liu Y, Whitman WB (2008) Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. Ann N Y Acad Sci 1125:171–189

Liu Y, Zhou Z, Pan J, Baker BJ, Gu JD, Li M (2018) Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. ISME J 12:1021–1031

Londei P, Ferreira-Cerca S (2021) Ribosome biogenesis in Archaea. Front Microbiol 12:686977

Lopez-Garcia P, Moreira D (2020) The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol 5:655–667

Lopez-Garcia P, Moreira D (2023) The symbiotic origin of the eukaryotic cell. C R Biol 346:55–73

Ludt K, Soppa J (2019) Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. Biochem Soc Trans 47:933–944

Lundgren M, Malandrin L, Eriksson S, Huber H, Bernander R (2008) Cell cycle characteristics of crenarchaeota: unity among diversity. J Bacteriol 190:5362–5367

Ma J, Campbell A, Karlin S (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733–5745

MacLeod FI, von Kügelgen A, Lechowska MK, Parham J, Richard IA, Aguilar-Pine EJ, Burns BP, Baker BJ, Bharat TAM, Baum B (2025) An Asgard archaeon with internal membrane compartments. preprint at: https://doi.org/10.1101/2025.07.22.663070

Majernik AI, Lundgren M, McDermott P, Bernander R, Chong JP (2005) DNA content and nucleoid distribution in *Methanothermobacter thermautotrophicus*. J Bacteriol 187:1856–1858

Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P et al (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18:67–83

Malandrin L, Huber H, Bernander R (1999) Nucleoid structure and partition in *Methanococcus jannaschii*: an archaeon with multiple copies of the chromosome. Genetics 152:1315–1323

Manoharan L, Kozlowski JA, Murdoch RW, Loffler FE, Sousa FL, Schleper C (2019) Metagenomes from coastal marine sediments give insights into the ecological role and cellular features of Loki- and Thorarchaeota. mBio 10:10-1128

Margulis L, Sagan D (1986) Microcosmos: four billion years of evolution from our microbial ancestors. University of California Press

Marquez V, Frohlich T, Armache JP, Sohmen D, Donhofer A, Mikolajka A, Berninghausen O, Thomm M, Beckmann R, Arnold GJ et al (2011) Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins. J Mol Biol 405:1215–1232

Martin W, Kowallik K (1999) Annotated English translation of Mereschkowsky's 1905 Paper 'Über Natur Und Ursprung Der chromatophoren imPflanzenreiche'. Eur J Phycol 34:287–295

Martin W, Muller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37–41

Mattiroli F, Bhattacharyya S, Dyer PN, White AE, Sandman K, Burkhart BW, Byrne KR, Lee T, Ahn NG, Santangelo TJ et al (2017) Structure of histone-based chromatin in Archaea. Science 357:609–612

 © The Author(s)

Melnikov N, Junglas B, Halbi G, Nachmias D, Zerbib E, Gueta N, Upcher A, Zalk R, Sachse C, Bernheim-Groswasser A et al (2025) The Asgard archaeal ESCRT-III system forms helical filaments and remodels eukaryotic-like membranes. EMBO J 44:665–681

Melnikov S, Ben-Shem A, Garreau de Loubresse N, Jenner L, Yusupova G, Yusupov M (2012) One core, two shells: bacterial and eukaryotic ribosomes. Nat Struct Mol Biol 19:560–567

Meng J, Xu J, Qin D, He Y, Xiao X, Wang F (2014) Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. ISME J 8:650–659

Mereschkowsky C (1910) Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis. Biologisches Centralblatt 30:353–367

Moreira D, Lopez-Garcia P (1998) Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol 47:517–530

Nachmias D, Melnikov N, Zorea A, Sharon M, Yemini R, De-Picchoto Y, Tsirkas I, Aharoni A, Frohn B, Schwille P et al (2023) Asgard ESCRT-III and VPS4 reveal conserved chromatin binding properties of the ESCRT machinery. ISME J 17:117–129

Nobs SJ, Johnson MD, Williams TJ, Meltzer J, Vázquez-Campos X, MacLeod FI, Rowell K, Pitt M, Paul B, Shepherd DC, et al (2025) An Asgard archaeon from a modern analog of ancient microbial mats. preprint at: https://www.biorxiv.org/content/10.1101/2025.07.22.663070v1

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, Chee GJ, Hattori M, Kanai A, Atomi H et al (2011) Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res 39:3204–3223

Nussbaum P, Gerstner M, Dingethal M, Erb C, Albers SV (2021) The archaeal protein SepF is essential for cell division in Haloferax volcanii. Nat Commun 12:3469

Obita T, Saksena S, Ghazi-Tabatabai S, Gill DJ, Perisic O, Emr SD, Williams RL (2007) Structural basis for selective recognition of ESCRT-III by the AAA ATPase Vps4. Nature 449:735–739

Offre P, Spang A, Schleper C (2013) Archaea in biogeochemical cycles. Annu Rev Microbiol 67:437–457

Olsen GJ, Woese CR (1997) Archaeal genomics: an overview. Cell 89:991–994

Pelve EA, Lindas AC, Martens-Habbena W, de la Torre JR, Stahl DA, Bernander R (2011) Cdv-based cell division and cell cycle organization in the thaumarchaeon *Nitrosopumilus maritimus*. Mol Microbiol 82:555–566

Pende N, Sogues A, Megrian D, Sartori-Rupp A, England P, Palabikyan H, Rittmann SKR, Grana M, Wehenkel AM, Alzari PM et al (2021) SepF is the FtsZ anchor in archaea, with features of an ancestral cell division system. Nat Commun 12:3214

Penev PI, Fakhretaha-Aval S, Patel VJ, Cannone JJ, Gutell RR, Petrov AS, Williams LD, Glass JB (2020) Supersized ribosomal RNA expansion segments in Asgard Archaea. Genome Biol Evol 12:1694–1710

Phung DK, Rinaldi D, Langendijk-Genevaux PS, Quentin Y, Carpousis AJ, Clouet-d'Orval B (2013) Archaeal beta-CASP ribonucleases of the aCPSF1 family are orthologs of the eukaryal CPSF-73 factor. Nucleic Acids Res 41:1091–1103

Prangishvili D, Bamford DH, Forterre P, Iranzo J, Koonin EV, Krupovic M (2017) The enigmatic archaeal virosphere. Nat Rev Microbiol 15:724–739

Probst AJ, Auerbach AK, Moissl-Eichinger C (2013) Archaea on human skin. PLoS ONE 8:e65388

Radler P, Viehboeck T, Luo Z-H, Maslac N, Schmidt K, Hauschild R, Nobu MK, Bulgheresi S, Stradal TEB, Rottner K et al (2025) Dynamic protrusions mediate unique crawling motility in Asgard Archaea (Promethearchaeota). Preprint at https://www.biorxiv.org/content/10.1101/2025.11.30.690169v1

Rambo IM, Langwig MV, Leao P, De Anda V, Baker BJ (2022) Genomes of six viruses that infect Asgard archaea from deep-sea sediments. Nat Microbiol 7:953–961

Ranawat HM, Cajili MK, Lopez-Barbosa N, Quail T, Dame RT, Dodonova SO (2025) Cryo-EM reveals open and closed Asgard chromatin assemblies. Mol Cell 85:4152–4165 e4155

Reiter WD, Hudepohl U, Zillig W (1990) Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. Proc Natl Acad Sci USA 87:9509–9513

Rhind N, Russell P (1998) The Schizosaccharomyces pombe S-phase checkpoint differentiates between different types of DNA damage. Genetics 149:1729–1737

Rinke C, Chuvochina M, Mussig AJ, Chaumeil PA, Davin AA, Waite DW, Whitman WB, Parks DH, Hugenholtz P (2021) A standardized archaeal taxonomy for the Genome Taxonomy Database. Nat Microbiol 6:946–959

Rodrigues-Oliveira T, Wollweber F, Ponce-Toledo RI, Xu J, Rittmann SKR, Klingl A, Pilhofer M, Schleper C (2023) Actin cytoskeleton and complex cell architecture in an Asgard archaeon. Nature 613:332–339

Rodriguez RL, Tsementzi D, Luo C, Konstantinidis KT (2020) Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. Environ Microbiol 22:3394–3412

Rossi M, Rella R, Pensa M, Bartolucci S, De Rosa M, Gambacorta A, Raia CA, Dell'Aversano Orabona N (1986) Structure and properties of a thermophilic and thermostable DNA polymerase isolated from Sulfolobus solfataricus. Syst Appl Microbiol 7:337–341

Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14:255–274

Sakakibara N, Kelman LM, Kelman Z (2009) Unwinding the structure and function of the archaeal MCM helicase. Mol Microbiol 72:286–296

Samson RY, Abeyrathne PD, Bell SD (2016) Mechanism of archaeal MCM helicase recruitment to DNA replication origins. Molecular cell 61:287–296

Samson RY, Obita T, Freund SM, Williams RL, Bell SD (2008) A role for the ESCRT system in cell division in archaea. Science 322:1710–1713

Samson RY, Obita T, Hodgson B, Shaw MK, Chong PL, Williams RL, Bell SD (2011) Molecular and structural basis of ESCRT-III recruitment to membranes during archaeal cell division. Mol Cell 41:186–196

Sanders TJ, Lammers M, Marshall CJ, Walker JE, Lynch ER, Santangelo TJ (2019) TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. Mol Microbiol 111:784–797

Schleper C, Nicol GW (2010) Ammonia-oxidising archaea-physiology, ecology and evolution. Adv Micro Physiol 57:1–41

Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, Santarius U, Klenk HP, Zillig W (1995) Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. J Bacteriol 177:7050–7059

Schmitt E, Coureux PD, Kazan R, Bourgeois G, Lazennec-Schurdevin C, Mechulam Y (2020) Recent advances in archaeal translation initiation. Front Microbiol 11:584152

Schmitt E, Coureux PD, Monestier A, Dubiez E, Mechulam Y (2019) Start codon recognition in eukaryotic and archaeal translation initiation: a common structural core. Int J Mol Sci 20:939

Schoneberg J, Pavlin MR, Yan S, Righini M, Lee IH, Carlson LA, Bahrami AH, Goldman DH, Ren X, Hummer G et al (2018) ATP-dependent force generation and membrane scission by ESCRT-III and Vps4. Science 362:1423–1428

Seitz KW, Dombrowski N, Eme L, Spang A, Lombard J, Sieber JR, Teske AP, Ettema TJG, Baker BJ (2019) Asgard archaea capable of anaerobic hydrocarbon cycling. Nat Commun 10:1822

© The Author(s)

Seitz KW, Lazar CS, Hinrichs KU, Teske AP, Baker BJ (2016) Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. ISME J 10:1696–1705

Souza DP, Espadas J, Chaaban S, Moody ERR, Hatano T, Balasubramanian M, Williams TA, Roux A, Baum B (2025) Asgard archaea reveal the conserved principles of ESCRT-III membrane remodeling. Sci Adv 11:eads5255

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179

Spang A, Stairs CW, Dombrowski N, Eme L, Lombard J, Caceres EF, Greening C, Baker BJ, Ettema TJG (2019) Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. Nat Microbiol 4:1138–1148

Stachler AE, Turgeman-Grott I, Shtifman-Segal E, Allers T, Marchfelder A, Gophna U (2017) High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. Nucleic Acids Res 45:5208–5216

Stepanov VG, Fox GE (2021) Expansion segments in bacterial and archaeal 5S ribosomal RNAs. RNA 27:133–150

Stetter KO (2013) A brief history of the discovery of hyperthermophilic life. Biochem Soc Trans 41:416–420

Stevens KM, Swadling JB, Hocher A, Bang C, Gribaldo S, Schmitz RA, Warnecke T (2020) Histone variants in archaea and the evolution of combinatorial chromatin complexity. Proc Natl Acad Sci USA 117:33384–33395

Sun J, Evans PN, Gagen EJ, Woodcroft BJ, Hedlund BP, Woyke T, Hugenholtz P, Rinke C (2021) Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. ISME Commun 1:30

Sung S, Li F, Park YB, Kim JS, Kim AK, Song OK, Kim J, Che J, Lee SE, Cho Y (2014) DNA end recognition by the Mre11 nuclease dimer: insights into resection and repair of damaged DNA. EMBO J 33:2422–2435

Survery S, Hurtig F, Haq SR, Eriksson J, Guy L, Rosengren KJ, Lindas AC, Chi CN (2021) Heimdallarchaea encodes profilin with eukaryotic-like actin regulation and polyproline binding. Commun Biol 4:1024

Takai K, Horikoshi K (1999) Genetic diversity of archaea in deep-sea hydrothermal vent environments. Genetics 152:1285–1297

Takai K, Komatsu T, Inagaki F, Horikoshi K (2001) Distribution of archaea in a black smoker chimney structure. Appl Environ Microbiol 67:3618–3629

Takai K, Nakamura K, Toki T, Tsunogai U, Miyazaki M, Miyazaki J, Hirayama H, Nakagawa S, Nunoura T, Horikoshi K (2008) Cell proliferation at 122 degrees C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. Proc Natl Acad Sci USA 105:10949–10954

Takemata N, Samson RY, Bell SD (2019) Physical and functional compartmentalization of archaeal chromosomes. Cell 179:165–179.e118

Tarrason Risa G, Hurtig F, Bray S, Hafner AE, Harker-Kirschneck L, Faull P, Davis C, Papatziamou D, Mutavchiev DR, Fan C et al (2020) The proteasome controls ESCRT-III-mediated cell division in an archaeon. Science 369:eaaz2532

Tirumalai MR, Kaelber JT, Park DR, Tran Q, Fox GE (2020) Cryo-electron microscopy visualization of a large insertion in the 5S ribosomal RNA of the extremely halophilic archaeon *Halococcus morrhuae*. FEBS Open Bio 10:1938–1946

Tobiasson V, Luo J, Wolf YI, Koonin EV (2026) Dominant contribution of Asgard archaea to eukaryogenesis. Nature 650:141–149

Tornabene TG, Langworthy TA (1979) Diphytanyl and dibiphytanyl glycerol ether lipids of methanogenic archaebacteria. Science 203:51–53

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Valentine DL (2007) Adaptations to energy stress dictate the ecology and evolution of the Archaea. Nat Rev Microbiol 5:316–323

van Wolferen M, Pulschen AA, Baum B, Gribaldo S, Albers SV (2022) The cell biology of archaea. Nat Microbiol 7:1744–1755

Vetriani C, Jannasch HW, MacGregor BJ, Stahl DA, Reysenbach AL (1999) Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. Appl Environ Microbiol 65:4375–4384

Villanueva L, Schouten S, Damste JS (2017) Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the 'lipid divide'. Environ Microbiol 19:54–69

Wallin IE (1883) Symbionticism and the origin of species. The Williams & Wilkins Company, Baltimore

Wallin IE (1922) On the nature of mitochondria. I. Observations on mitochondria staining methods applied to bacteria. II. Reactions of bacteria to chemical treatment. Am J Anat 30:203–229

Ward HM (1883) Ueber die Entwickelung der Chlorophyllkörner und Farbkörper Ueber Chlorophyllkörner Stärkebildner und Farbkörper. Nature 28:267–268

Wenck BR, Santangelo TJ (2020) Archaeal transcription. Transcription 11:199–210

White MF (2011) Homologous recombination in the archaea: the means justify the ends. Biochem Soc Trans 39:15–19

White MF, Allers T (2018) DNA repair in the archaea—an emerging picture. FEMS Microbiol Rev 42:514–526

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. Proc Biol Sci 279:4870–4879

Williamson K, Eme L, Baños H, McCarthy CGP, Susko E, Kamikawa R, Orr RJS, Muñoz-Gómez SA, Minh BQ, Simpson AGB et al (2025) A robustly rooted tree of eukaryotes reveals their excavate ancestry. Nature 640:974–981

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR (2004) A new biology for a new century. Microbiol Mol Biol Rev 68:173–186

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Woese CR, Magrum LJ, Fox GE (1978) Archaebacteria. J Mol Evol 11:245–251

Wollweber F, Xu J, Ponce-Toledo RI, Marxer F, Rodrigues-Oliveira T, Possnecker A, Luo ZH, Malit JJL, Kokhanovska A, Wieczorek M et al (2025) Microtubules in Asgard archaea. Cell 188:2451–2464.e2426

Woodman IL, Bolt EL (2009) Molecular biology of Hel308 helicase in archaea. Biochem Soc Trans 37:74–78

Wu F, Speth DR, Philosof A, Cremiere A, Narayanan A, Barco RA, Connon SA, Amend JP, Antoshechkin IA, Orphan VJ (2022) Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularized Asgard archaea genomes. Nat Microbiol 7:200–212

Xie R, Wang Y, Huang D, Hou J, Li L, Hu H, Zhao X, Wang F (2022) Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. Sci China Life Sci 65:818–829

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origins. Proc Natl Acad Sci USA 82:4443–4447

Yin X, Cai M, Liu Y, Zhou G, Richter-Heitmann T, Aromokeye DA, Kulkarni AC, Nimzyk R, Cullhed H, Zhou Z et al (2021) Subgroup level differences of physiological activities in marine Lokiarchaeota. ISME J 15:848–861

Yutin N, Koonin EV (2012) Archaeal origin of tubulin. Biol Direct 7:10

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU et al

 © The Author(s)

(2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358

Zerulla K, Soppa J (2014) Polyploidy in haloarchaea: advantages for growth and survival. Front Microbiol 5:274

Zhang C, Phillips APR, Wipfler RL, Olsen GJ, Whitaker RJ (2018) The essential genome of the crenarchaeal model *Sulfolobus islandicus*. Nat Commun 9:4908

Zhang C, Whitaker RJ (2018) Microhomology-mediated high-throughput gene inactivation strategy for the hyperthermophilic crenarchaeon *Sulfolobus islandicus*. Appl Environ Microbiol 84:e02167-17

Zhang CJ, Chen YL, Sun YH, Pan J, Cai MW, Li M (2021) Diversity, metabolism and cultivation of archaea in mangrove ecosystems. Mar Life Sci Technol 3:252–262

Zhang J, Feng X, Li M, Liu Y, Liu M, Hou LJ, Dong HP (2025) Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota. Nature 642:990–998

Zillig W, Palm P, Reiter WD, Gropp F, Puhler G, Klenk HP (1988) Comparative evaluation of gene expression in archaebacteria. Eur J Biochem 173:473–482

Zillig W, Schnabel R, Stetter KO (1985) Archaebacteria and the origin of the eukaryotic cytoplasm. Curr Top Microbiol Immunol 114:1–18

Zillig W, Stetter KO, Janekovic D (1979) DNA-dependent RNA polymerase from the archaebacterium *Sulfolobus acidocaldarius*. Eur J Biochem 96:597–604

Zink IA, Wimmer E, Schleper C (2020) Heavily armed ancestors: CRISPR immunity and applications in Archaea with a comparative analysis of CRISPR types in Sulfolobales. Biomolecules 10:1523

Zou D, Liu H, Li M (2020) Community, distribution, and ecological roles of estuarine archaea. Front Microbiol 11:2060

Acknowledgements

We would like to thank all colleagues who have contributed to the archaeal field over the years, for their discussions and interactions. We apologize for not having cited everyone. We thank Frederic Berger, Nevena Maslac, Philipp Radler as well as the four reviewers for their input and important comments on the manuscript. We are also indebted to Nathalia Jandl for help with formatting references. Thanks to Kai Finster for pointing out the Darwin citation. This work was supported by the Austrian Science Fund (FWF) with W1257 (Wittgenstein) and EFP 25 (Emerging Field Project) to CS.

Author contributions

Christa Schleper: Conceptualization; Funding acquisition; Writing—original draft; Writing—review and editing. **Thiago Rodrigues-Oliveira**: Writing—original draft; Writing—review and editing.

Disclosure and competing interests statement

The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the data associated with this article, unless otherwise stated in a credit line to the data, but does not extend to the graphical or creative elements of illustrations, charts, or figures. This waiver removes legal barriers to the re-use and mining of research data. According to standard scholarly practice, it is recommended to provide appropriate citation and attribution whenever technically possible.

© The Author(s) 2026

D.5 Genomics

Further explanations for Chapter 5.

D.5.1 Sequencing methods

- Heather and Chain (2016) — A historical review of DNA sequencing from Sanger and Maxam–Gilbert through second-generation platforms to single-molecule methods. Useful for placing the three generations of the section in context. *(review; CC BY 4.0)*

  The sequence of sequencers: The history of sequencing DNA. Heather & Chain (2016). *Genomics*. doi:10.1016/j.ygeno.2015.11.003

  > Determining the order of nucleic acid residues in biological samples is an integral component of a wide variety of research applications. Over the last fifty years large numbers of researchers have applied themselves to the production of techniques and technologies to facilitate this feat, sequencing DNA and RNA molecules. This time-scale has witnessed tremendous changes, moving from sequencing short oligonucleotides to millions of bases, from struggling towards the deduction of the coding sequence of a single gene to rapid and widely available whole genome sequencing. This article traverses those years, iterating through the different generations of sequencing technology, highlighting some of the key discoveries, researchers, and sequences along the way.

- Satam et al. (2023) — A broad overview of current short- and long-read sequencing platforms, their chemistry, read lengths and error profiles, with an introduction to typical applications. *(review; CC BY 4.0)*

  Next-Generation Sequencing Technology: Current Trends and Advancements. Satam et al. (2023). *Biology*. doi:10.3390/biology12070997

  > The advent of next-generation sequencing (NGS) has brought about a paradigm shift in genomics research, offering unparalleled capabilities for analyzing DNA and RNA molecules in a high-throughput and cost-effective manner. This transformative technology has swiftly propelled genomics advancements across diverse domains. NGS allows for the rapid sequencing of millions of DNA fragments simultaneously, providing comprehensive insights into genome structure, genetic variations, gene expression profiles, and epigenetic modifications. The versatility of NGS platforms has expanded the scope of genomics research, facilitating studies on rare genetic diseases, cancer genomics, microbiome analysis, infectious diseases, and population genetics. Moreover, NGS has enabled the development of targeted therapies, precision medicine approaches, and improved diagnostic methods. This review provides an insightful overview of the current trends and recent advancements in NGS technology, highlighting its potential impact on diverse areas of genomic research. Moreover, the review delves into the challenges encountered and future directions of NGS technology, including endeavors to enhance the accuracy and sensitivity of sequencing data, the development of novel algorithms for data analysis, and the pursuit of more efficient, scalable, and cost-effective solutions that lie ahead.

- Amarasinghe et al. (2020) — Reviews PacBio and Oxford Nanopore technologies and the computational steps that follow, such as base calling, error correction and alignment of long reads. Complements the section on third-generation sequencing. *(review; CC BY 4.0)*

  Opportunities and challenges in long-read sequencing data analysis. Amarasinghe et al. (2020). *Genome Biology*. doi:10.1186/s13059-020-1935-5

  > Long-read technologies are overcoming early limitations in accuracy and throughput, broadening their application domains in genomics. Dedicated analysis tools that take into account the characteristics of long-read data are thus required, but the fast pace of development of such tools can be overwhelming. To assist in the design and analysis of long-read sequencing projects, we review the current landscape of available tools and present an online interactive database, long-read-tools.org, to facilitate their browsing. We further focus on the principles of error correction, base modification detection, and long-read transcriptomics analysis and highlight the challenges that remain.

D.5.2 Genome assembly

- Jung et al. (2020) — A short, practical guide to assembling and annotating a genome, from choosing a sequencing strategy to quality assessment. Gives a hands-on view of the workflow described in this section and in the gene prediction section. *(tutorial; CC BY 4.0)*

  Twelve quick steps for genome assembly and annotation in the classroom. Jung et al. (2020). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1008325

  > Eukaryotic genome sequencing and de novo assembly, once the exclusive domain of well-funded international consortia, have become increasingly affordable, thus fitting the budgets of individual research groups. Third-generation long-read DNA sequencing technologies are increasingly used, providing extensive genomic toolkits that were once reserved for a few select model organisms. Generating high-quality genome assemblies and annotations for many aquatic species still presents significant challenges due to their large genome sizes, complexity, and high chromosome numbers. Indeed, selecting the most appropriate sequencing and software platforms and annotation pipelines for a new genome project can be daunting because tools often only work in limited contexts. In genomics, generating a high-quality genome assembly/annotation has become an indispensable tool for better understanding the biology of any species. Herein, we state 12 steps to help researchers get started in genome projects by presenting guidelines that are broadly applicable (to any species), sustainable over time, and cover all aspects of genome assembly and annotation projects from start to finish. We review some commonly used approaches, including practical methods to extract high-quality DNA and choices for the best sequencing platforms and library preparations. In addition, we discuss the range of potential bioinformatics pipelines, including structural and functional annotations (e.g., transposable elements and repetitive sequences). This paper also includes information on how to build a wide community for a genome project, the importance of data management, and how to make the data and results Findable, Accessible, Interoperable, and Reusable (FAIR) by submitting them to a public repository and sharing them with the research community.

- Dominguez Del Angel et al. (2018) — An introductory guide covering planning, data types, assembly, quality assessment and annotation of eukaryotic genomes. Useful for understanding the practical choices behind assembly projects. *(tutorial; CC BY 4.0)*

  Ten steps to get started in Genome Assembly and Annotation. Dominguez Del Angel et al. (2018). *F1000Research*. doi:10.12688/f1000research.13598.1

  > As a part of the ELIXIR-EXCELERATE efforts in capacity building, we present here 10 steps to facilitate researchers getting started in genome assembly and genome annotation. The guidelines given are broadly applicable, intended to be stable over time, and cover all aspects from start to finish of a general assembly and annotation project. Intrinsic properties of genomes are discussed, as is the importance of using high quality DNA. Different sequencing technologies and generally applicable workflows for genome assembly are also detailed. We cover structural and functional annotation and encourage readers to also annotate transposable elements, something that is often omitted from annotation workflows. The importance of data management is stressed, and we give advice on where to submit data and how to make your results Findable, Accessible, Interoperable, and Reusable (FAIR).

- Ekblom and Wolf (2014) — A readable guide for biologists to sequencing strategies, de Bruijn graph and overlap assembly, scaffolding, assembly quality metrics and annotation. Some technology details are dated, but the concepts match this chapter well. *(review; CC BY 3.0)*

  A field guide to whole-genome sequencing, assembly and annotation. Ekblom & Wolf (2014). *Evolutionary Applications*. doi:10.1111/eva.12178

  > Genome sequencing projects were long confined to biomedical model organisms and required the concerted effort of large consortia. Rapid progress in high-throughput sequencing technology and the simultaneous development of bioinformatic tools have democratized the field. It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate de novo draft genome sequences for any organism of choice. Because of the cost and considerable effort involved in such an endeavour, the important first step is to thoroughly consider whether a genome sequence is necessary for addressing the biological question at hand. Once this decision is taken, a genome project requires careful planning with respect to the organism involved and the intended quality of the genome draft. Here, we briefly review the state of the art within this field and provide a step-by-step introduction to the

workflow involved in genome sequencing, assembly and annotation with particular reference to large and complex genomes. This tutorial is targeted at scientists with a background in conservation genetics, but more generally, provides useful practical guidance for researchers engaging in whole-genome sequencing projects.

D.5.3 Mapping and variant calling

- Alser et al. (2021) — Surveys read-mapping algorithms, including indexing with hash tables, the Burrows–Wheeler transform and minimizers, seeding, filtering and alignment, and how they have followed sequencing technology. Explains the algorithms behind the read-mapping subsection. *(review; CC BY 4.0)*

  Technology dictates algorithms: recent developments in read alignment. Alser et al. (2021). *Genome Biology*. doi:10.1186/s13059-021-02443-7

  > Aligning sequencing reads onto a reference is an essential step of the majority of genomic analysis pipelines. Computational algorithms for read alignment have evolved in accordance with technological advances, leading to today's diverse array of alignment methods. We provide a systematic survey of algorithmic foundations and methodologies across 107 alignment methods, for both short and long reads. We provide a rigorous experimental evaluation of 11 read aligners to demonstrate the effect of these underlying algorithms on speed and efficiency of read alignment. We discuss how general alignment algorithms have been tailored to the specific needs of various domains in biology.

- Zverinova and Guryev (2021) — Reviews the steps from aligned reads to called variants, covering small variants and structural variants, sources of error and benchmarking. A good companion to the variant-calling subsection. *(review; CC BY 4.0)*

  Variant calling: Considerations, practices, and developments. Zverinova & Guryev (2021). *Human Mutation*. doi:10.1002/humu.24311

  > The success of many clinical, association, or population genetics studies critically relies on properly performed variant calling step. The variety of modern genomics protocols, techniques, and platforms makes our choices of methods and algorithms difficult and there is no ”one size fits all” solution for study design and data analysis. In this review, we discuss considerations that need to be taken into account while designing the study and preparing for the experiments. We outline the variety of variant types that can be detected using sequencing approaches and highlight some specific requirements and basic principles of their detection. Finally, we cover interesting developments that enable variant calling for a broad range of applications in the genomics field. We conclude by discussing technological and algorithmic advances that have the potential to change the ways of calling DNA variants in the nearest future.

- Koboldt (2020) — Describes a practical variant-calling workflow (alignment, duplicate marking, base quality recalibration, calling and filtering) for germline and somatic variants. Clinically oriented, but the workflow is general. *(review; CC BY 4.0)*

  Best practices for variant calling in clinical sequencing. Koboldt (2020). *Genome Medicine*. doi:10.1186/s13073-020-00791-w

  > Next-generation sequencing technologies have enabled a dramatic expansion of clinical genetic testing both for inherited conditions and diseases such as cancer. Accurate variant calling in NGS data is a critical step upon which virtually all downstream analysis and interpretation processes rely. Just as NGS technologies have evolved considerably over the past 10 years, so too have the software tools and approaches for detecting sequence variants in clinical samples. In this review, I discuss the current best practices for variant calling in clinical sequencing studies, with a particular emphasis on trio sequencing for inherited disorders and somatic mutation detection in cancer patients. I describe the relative strengths and weaknesses of panel, exome, and whole-genome sequencing for variant detection. Recommended tools and strategies for calling variants of different classes are also provided, along with guidance on variant review, validation, and benchmarking to ensure optimal performance. Although NGS technologies are continually evolving, and new capabilities (such as long-read single-molecule sequencing) are emerging, the ”best practice” principles in this review should be relevant to clinical variant calling in the long term.

D.5.4 Gene prediction

- Ejigu and Jung (2020) — Reviews structural and functional genome annotation, including ab initio gene finders, evidence-based and combined pipelines, for prokaryotes and eukaryotes. Matches the scope of this section. *(review; CC BY 4.0)*

  Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Ejigu & Jung (2020). *Biology*. doi:10.3390/biology9090295

  > Next-Generation Sequencing (NGS) has made it easier to obtain genome-wide sequence data and it has shifted the research focus into genome annotation. The challenging tasks involved in annotation rely on the currently available tools and techniques to decode the information contained in nucleotide sequences. This information will improve our understanding of general aspects of life and evolution and improve our ability to diagnose genetic disorders. Here, we present a summary of both structural and functional annotations, as well as the associated comparative annotation tools and pipelines. We highlight visualization tools that immensely aid the annotation process and the contributions of the scientific community to the annotation. Further, we discuss quality-control practices and the need for re-annotation, and highlight the future of annotation.

- Salzberg (2019) — A short commentary on why gene annotation of new genomes remains error-prone, for example through contamination and incomplete evidence. Useful for discussing the limits of automated annotation. *(perspective; CC BY 4.0)*

  Next-generation genome annotation: we still struggle to get it right. Salzberg (2019). *Genome Biology*. doi:10.1186/s13059-019-1715-2

  > While the genome sequencing revolution has led to the sequencing and assembly of many thousands of new genomes, genome annotation still uses very nearly the same technology that we have used for the past two decades. The sheer number of genomes necessitates the use of fully automated procedures for annotation, but errors in annotation are just as prevalent as they were in the past, if not more so. How are we to solve this growing problem?

D.5.5 Metagenomics

- Wooley et al. (2010) — An introduction to metagenomics for computational biologists, covering sampling, 16S studies, shotgun sequencing, assembly, binning and functional analysis. Older, but explains the basic concepts clearly. *(primer; CC BY 4.0)*

  A Primer on Metagenomics. Wooley et al. (2010). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000667

  > Metagenomics is a discipline that enables the genomic study of uncultured microorganisms. Faster, cheaper sequencing technologies and the ability to sequence uncultured microbes sampled directly from their habitats are expanding and transforming our view of the microbial world. Distilling meaningful information from the millions of new genomic sequences presents a serious challenge to bioinformaticians. In cultured microbes, the genomic data come from a single clone, making sequence assembly and annotation tractable. In metagenomics, the data come from heterogeneous microbial communities, sometimes containing more than 10,000 species, with the sequence data being noisy and partial. From sampling, to assembly, to gene calling and function prediction, bioinformatics faces new demands in interpreting voluminous, noisy, and often partial sequence data. Although metagenomics is a relative newcomer to science, the past few years have seen an explosion in computational methods applied to metagenomic-based research. It is therefore not within the scope of this article to provide an exhaustive review. Rather, we provide here a concise yet comprehensive introduction to the current computational requirements presented by metagenomics, and review the recent progress made. We also note whether there is software that implements any of the methods presented here, and briefly review its utility. Nevertheless, it would be useful if readers of this article would avail themselves of the comment section provided by this journal, and relate their own experiences. Finally, the last section of this article provides a few representative studies illustrating different facets of recent scientific discoveries made using metagenomics.

- Pérez-Cobas et al. (2020) — Compares amplicon (marker gene) and shotgun metagenomics and describes the analysis steps and tools for each. Covers the amplicon and shotgun subsections. *(review; CC BY 4.0)*

  Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Pérez-Cobas et al. (2020). *Microbial Genomics*. doi:10.1099/mgen.0.000409

  > Metagenomics and marker gene approaches, coupled with high-throughput sequencing technologies, have revolutionized the field of microbial ecology. Metagenomics is a culture-independent method that allows the identification and characterization of organisms from all kinds of samples. Whole-genome shotgun sequencing analyses the total DNA of a chosen sample to determine the presence of micro-organisms from all domains of life and their genomic content. Importantly, the whole-genome shotgun sequencing approach reveals the genomic diversity present, but can also give insights into the functional potential of the micro-organisms identified. The marker gene approach is based on the sequencing of a specific gene region. It allows one to describe the microbial composition based on the taxonomic groups present in the sample. It is frequently used to analyse the biodiversity of microbial ecosystems. Despite its importance, the analysis of metagenomic sequencing and marker gene data is quite a challenge. Here we review the primary workflows and software used for both approaches and discuss the current challenges in the field.

- Lapidus and Korobeynikov (2021) — Reviews why metagenome assembly is harder than single-genome assembly and how metagenomic assemblers and binning lead to metagenome-assembled genomes. *(review; CC BY 4.0)*

  Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. Lapidus & Korobeynikov (2021). *Frontiers in Microbiology*. doi:10.3389/fmicb.2021.613791

  > Metagenomics is a segment of conventional microbial genomics dedicated to the sequencing and analysis of combined genomic DNA of entire environmental samples. The most critical step of the metagenomic data analysis is the reconstruction of individual genes and genomes of the microorganisms in the communities using metagenomic assemblers - computational programs that put together small fragments of sequenced DNA generated by sequencing instruments. Here, we describe the challenges of metagenomic assembly, a wide spectrum of applications in which metagenomic assemblies were used to better understand the ecology and evolution of microbial ecosystems, and present one of the most efficient microbial assemblers, SPAdes that was upgraded to become applicable for metagenomics.

D.5.6 Regulatory and epigenomic sequencing data

- Ma and Zhang (2020) — An open review of how ChIP-seq and ATAC-seq work and how their data are analysed, from peak calling to integration with other data. *(review; CC BY 4.0)*

  Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. Ma & Zhang (2020). *Molecular Biomedicine*. doi:10.1186/s43556-020-00009-w

  > Chromatin regulatory landscape plays a critical role in many disease processes and embryo development. Epigenome sequencing technologies such as chromatin immunoprecipitation sequencing (ChIP-seq) and assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) have enabled us to dissect the pan-genomic regulatory landscape of cells and tissues in both time and space dimensions by detecting specific chromatin state and its corresponding transcription factors. Pioneered by the advancement of chromatin immunoprecipitation-chip (ChIP-chip) technology, abundant epigenome profiling technologies have become available such as ChIP-seq, DNase I hypersensitive site sequencing (DNase-seq), ATAC-seq and so on. The advent of single-cell sequencing has revolutionized the next-generation sequencing, applications in single-cell epigenetics are enriched rapidly. Epigenome sequencing technologies have evolved from low-throughput to high-throughput and from bulk sample to the single-cell scope, which unprecedentedly benefits scientists to interpret life from different angles. In this review, after briefly introducing the background knowledge of epigenome biology, we discuss the development of epigenome sequencing technologies, especially ChIP-seq & ATAC-seq and their current applications in scientific research. Finally, we provide insights into future applications and challenges.

D.5.7 Human genetics: finding disease genes and variants

- Visscher et al. (2017) — A balanced account of what genome-wide association studies have and have not achieved, including heritability, polygenicity and translation. *(review; free to read (Elsevier open archive))*

  10 Years of GWAS Discovery: Biology, Function, and Translation. Visscher et al. (2017). *The American Journal of Human Genetics*. doi:10.1016/j.ajhg.2017.06.005

  > Application of the experimental design of genome-wide association studies (GWASs) is now 10 years old (young), and here we review the remarkable range of discoveries it has facilitated in population and complex-trait genetics, the biology of diseases, and translation toward new therapeutics. We predict the likely discoveries in the next 10 years, when GWASs will be based on millions of samples with array data imputed to a large fully sequenced reference panel and on hundreds of thousands of samples with whole-genome sequencing data.

The following open-access paper(s) are reproduced under their original licences.

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Twelve quick steps for genome assembly and annotation in the classroom

Hyungtaek Jung[1,2]*, Tomer Ventura[3], J. Sook Chung[4], Woo-Jin Kim[5], Bo-Hye Nam[6], Hee Jeong Kong[6], Young-Ok Kim[6], Min-Seung Jeon[7], Seong-il Eyun[7]*

1 School of Biological Sciences, The University of Queensland, St Lucia, Queensland, Australia, **2** Centre for Agriculture and Bioeconomy, Queensland University of Technology, Brisbane, Queensland, Australia, **3** Genecology Research Centre, School of Science and Engineering, University of the Sunshine Coast, Sippy Downs, Queensland, Australia, **4** Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Baltimore, Maryland, United States of America, **5** Genetics and Breeding Research Center, National Institute of Fisheries Science, Geoje, Korea, **6** Biotechnology Research Division, National Institute of Fisheries Science, Busan, Korea, **7** Department of Life Science, Chung-Ang University, Seoul, Korea

* hyungtaek.jung@uq.edu.au (HJ); eyun@cau.ac.kr (SE)

Check for updates

OPEN ACCESS

Citation: Jung H, Ventura T, Chung JS, Kim W-J, Nam B-H, Kong HJ, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLoS Comput Biol 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Editor: Francis Ouellette, University of Toronto, CANADA

Published: November 12, 2020

Copyright: © 2020 Jung et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Korean Ministry of Agriculture, Food, and Rural Affairs (918010042HD030, Strategic Initiative for Microbiomes in Agriculture and Food) to SE. This work was also supported by a grant from the National Institute of Fisheries Science (R2020001) to WK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Eukaryotic genome sequencing and de novo assembly, once the exclusive domain of well-funded international consortia, have become increasingly affordable, thus fitting the budgets of individual research groups. Third-generation long-read DNA sequencing technologies are increasingly used, providing extensive genomic toolkits that were once reserved for a few select model organisms. Generating high-quality genome assemblies and annotations for many aquatic species still presents significant challenges due to their large genome sizes, complexity, and high chromosome numbers. Indeed, selecting the most appropriate sequencing and software platforms and annotation pipelines for a new genome project can be daunting because tools often only work in limited contexts. In genomics, generating a high-quality genome assembly/annotation has become an indispensable tool for better understanding the biology of any species. Herein, we state 12 steps to help researchers get started in genome projects by presenting guidelines that are broadly applicable (to any species), sustainable over time, and cover all aspects of genome assembly and annotation projects from start to finish. We review some commonly used approaches, including practical methods to extract high-quality DNA and choices for the best sequencing platforms and library preparations. In addition, we discuss the range of potential bioinformatics pipelines, including structural and functional annotations (e.g., transposable elements and repetitive sequences). This paper also includes information on how to build a wide community for a genome project, the importance of data management, and how to make the data and results Findable, Accessible, Interoperable, and Reusable (FAIR) by submitting them to a public repository and sharing them with the research community.

Introduction

Genome projects employ state-of-the-art DNA sequencing, mapping, and computational technologies (including cross-disciplinary experimental designs) to expand our knowledge and

understanding of molecular/cellular mechanisms, gene repertoires, genome architecture, and evolution. The revolution in new sequencing technologies and computational developments has allowed researchers to drive advances in genome assembly and annotation to make the process better, faster, and cheaper with key model organisms [1,2].

Such technical advantages and established recommendations and strategies have been widely applied in humans [3–6], terrestrial animals [7–12], and plants and crops [13–18]. Genomic applications in aquatic species that could be potentially important for aquaculture are slower compared with human, livestock, and crops [19–21], compounded by larger diversity, lack of reference genomes, and more novice aquaculture industries. Given that aquaculture is the most rapidly expanding food sector, with the widest diversity of species cultured, it is poised for rapid adoption of genomics applications as these become more accessible. For any specific advice on application of genomics to aquaculture, please refer to previous works [19–25].

Before genome sequencing, a must-have step involves RNA sequencing (RNA-seq) that has provided significant insights into the biological functions [26–30]. RNA-seq plays a key role in genome annotation [31–36] through the identification of protein-coding genes based on transcriptome sequencing data and ab initio or homology-based prediction. However, the use of RNA-seq for genome assembly is limited to genome scaffolding [37]. While RNA-seq is a powerful technology that will likely remain a key asset in the biologist's toolkit, recent single-molecule mRNA sequencing approaches (e.g., Pacific Bioscience [PacBio] and Oxford Nanopore Technology [ONT]) have provided significant improvements in gene and genome annotation, making them appealing alternatives or complementary techniques for genome annotation [38–40].

Restriction site–associated DNA sequencing and diversity array technology are cost-effective methods that mainly focus on the detection of loci and the segregation of variants or genome-wide single nucleotide polymorphisms. The generation of genetic linkage maps has been successfully applied to recognize key components in the sustainable production of aquaculture species [41,42]. These attempts have resulted in the emphasis of genomic evaluations/selections or advanced selective breeding programs for desirable traits, such as growth, sex determination, sex markers, and disease resistance [42]. While these inexpensive techniques have been powerful tools for understanding the genetics of adaptation, recent studies have indicated their limitations for genome scans because they will likely miss many loci under selection, particularly for species with short linkage disequilibrium [43]. However, the widespread use of whole-genome sequencing (WGS) allows the detection of a full range of common and rare/hidden genetic variants of different types across almost entire genomes.

Many seminal biological discoveries in the 20th century were made using only a genetic analysis of a few selected model organisms because they were readily available for genetic analysis [44]. However, a high-quality and well-annotated genome assembly is increasingly becoming an essential tool for applied and basic research across many biological disciplines in the 21st century that can turn any organism into a model organism. Thus, securing more complete and accurate reference genomes and annotations before analyzing post-genome studies such as genome-wide association studies, structural variations, and posttranslational studies (methylation or histone modification) has become a cornerstone of modern genomics. Chromosome-level high-quality genomes (including structural and functional annotations) are differentiated from draft genomes by their completeness (low number of gaps and ambiguous *N*s), low number of assembly errors, and a high percentage of sequences assembled into chromosomes. Advances in next-generation sequencing (NGS) technologies and their analytical tools have made assembling and annotating the genomic sequence of most organisms both more feasible and affordable [33,45,46]. Table 1 shows recent chromosome-level genome

Table 1. Summary of recently published chromosome-level genome assemblies in aquaculture species using long-read sequences[a,b].

Scientific name	GS (Gb)	Final output			Input detail and depth (×)					BAs	Reference
		AGS (Gb)	FSN	N50 (Mb)	IM	PacBio	ONT	10xGC	Hi-C (×)		
Fish											
Collichthys lucidus	0.83/DP	0.88	24	1.1	63	109			233	Sex determination genes and chromosomes	[72]
Clupea harengus	0.81/DP	0.79	26	29.85	50	76			20	Chromosome rearrangement and spawning time	[10]
Epinephelus akaara	1.11/DP	1.14	24	46.03	49		96		100	Chromosome-level reference genome	[73]
Epinephelus lanceolatus	1.07/DP	1.09	24	46.2	134				0.2	Innate immunity and growth	[74]
Sebastes schlegelii	0.87/DP	0.81	24	3.85	132	66			189	Maternal reproductive system	[75]
Lateolabrax maculatus	0.65/DP	0.67	24	22.34	321				109	Chromosome-level reference genome	[41]
Oplegnathus fasciatus	0.78/DP	0.77	24	33.5	116	80			118	Chromosome-level reference genome	[43]
Pelteobagrus fulvidraco	0.72/DP	0.73	26	25.8	70	53			200	Chromosome-level reference genome	[40]
Shellfish											
Sinonovacula constricta	1.33/DP	1.22	19	65.93	148	148	136	123	154	Chromosome-level reference genome	[76]

[a]This table represents a selection of recent aquaculture genome works focusing on whole-genome assemblies using BioNano and/or Hi-C data (at least 1 technology used) since 2018. In addition, the table does not include any pure TGS/SGS/hybrid genome assemblies without BioNano/Hi-C data, single-cell sequencing, or transcriptomes. If the original report had no estimated input depth, this was calculated from the raw data. For the most recent global statistics, we highly recommend visiting the associated GenBank BioProject.

[b]AGS, assembled genome size; BAs, biological applications; DP, diploid; FSN, final scaffold number (pseudochromosome number); GS, genome size; IM, Illumina (combined paired-end [PE] and mate-pair [MP] reads); ONT, Oxford Nanopore Technology; PacBio, Pacific Bioscience; SGS, second-generation sequencing; TGS, third-generation sequencing; 10xGC, 10x Genomics Chromium.

https://doi.org/10.1371/journal.pcbi.1008325.t001

assemblies and provides a rough estimate of the sequencing depth and costs for beginners to achieve a chromosome-level genome assembly. For diploids, using a minimum 60× depth for PacBio, ONT, 60× for Illumina (San Diego, California, United States of America), and 100× for Hi-C data (Phase Genomics, Seattle, Washington, USA) (an extension of chromosome conformation capture, 3C) is recommended. High-quality end-to-end genome assembly and annotation of small eukaryotic (approximately 1 Gb diploid) and prokaryotic organisms have been achievable with small-to-medium financial resources and limited time, labor, and skill commitments. Nearly all eukaryotic genomes still represent a significant challenge for most aquatic species that have large and complex genomes and no reference genomes.

Furthermore, the following fundamental questions should be addressed: Why are genome projects or WGS necessary? What is the aim of a genome project? What kind of information is the research community expected to gather? Even from the beginning of a genome project, describing the expected end product, including project duration/budget, chromosome end-to-end completion, genome browser, and research paper, is required. In particular, if budget is a major obstacle, the best option to raise funds to support the genome project (e.g., industry or government support) must be determined. In addition to the abovementioned limitations, another essential element is bioinformatics, which has become a common denominator to produce and use software that can be applied to biological data in different contexts. As big data and multi-omics analyses are becoming mainstream, computational proficiency and literacy are indispensable skills in a biologist's toolkit in modern scientific society. All "omics" studies require a certain degree of computational biology: The implementation of analyses requires programming skills and knowledge of computer languages, while experimental design and interpretation require a solid understanding of analytical approaches [47,48]. These could

be daunting tasks for biologists who are unfamiliar with computational standards (e.g., codes, pipelines, and system environments) and resources (e.g., SourceForge, Bitbucket, GitLab, and GitHub). While academic cores, commercial services, and collaborations can aid in the implementation of analyses, the computational literacy required to design and interpret omics studies cannot simply be replaced or supplemented [47,48].

In the absence of a standard approach for genome projects, this paper aims to provide practical steps to facilitate project completion before embarking upon a genome assembly and annotation project (mainly for eukaryotic genomes). The target audience is anyone entering this field for the first time, particularly those who do not specialize in genomics research. While we can strive to answer questions in a manner that considers the beginner's perspective, certain aspects (e.g., assembly algorithms and computer environments) might require further reading for an in-depth understanding.

Step 1: Build a wide community for the project if possible

All genome projects have a common but monumental goal: sequencing the entire target genome for a wide range of genomics applications. While genomics is a rich field, one of the most prominent scientific objectives is probably securing the future of sustainable food sources by harnessing the power of genomics (i.e., desirable traits) [19–21,23–25], particularly for agriculture. If the species of interest is distinct from the wild, cultured, or harvested, it necessitates networking and building a scientific or stakeholder community to support the project. This usually requires a multi-institutional effort to both initiate and—more importantly—complete the genome project and then interpret the vast quantities of sequencing information produced for any given organism. As expected, WGS/genome projects' infrastructure demands are particularly high as varying interpretations may require facilities, personnel (skill intensive), and software (knowledge intensive) that suit the needs of immediate analyses, ongoing reanalyses, and the integration of genomic and other phenotype information (or desirable traits). Data storage, maintenance, transfer, and analysis costs will also likely remain substantial and represent an increasing proportion of overall sequencing costs in the future. Moreover, professional groups (including students), expert panels, and field farmers acknowledge that there is a need for educational programs specific to WGS demands. Addressing these needs will likely require substantial investment by agriculture production care systems. Thus, the real cost of WGS—including ongoing maintenance—could be even higher. Despite these burdens, most genome projects bring together leading researchers to work together and build large datasets of DNA from target genomes, which has significantly benefited the research community. These efforts facilitate the sharing of sequence data and help research advance. In particular, smaller research groups that have less experience and are poorly equipped in areas including raw read sequencing and assembly and annotation should consider the main features and steps outlined here via community collaboration. In the case of funding for genome projects, applying for government grants and receiving corporate sponsorships as a consortium could be considered potential solutions as these avenues have been successful for humans, livestock (cow, pig, and sheep), crops (Arabidopsis, rice, and tomato), and aquaculture (salmon, oyster tilapia, and prawn).

Step 2: Gather information about the target genome

Every genome sequencing, assembly, and annotation project is different due to each subject genome's distinctive properties. There are four fundamental aspects that must be considered when embarking on a new genome project: the genome size, levels of ploidy and heterozygosity, GC content, and complexity. These will directly affect the overall quality and cost of genome sequencing, assembly, and annotation [14,49].

How big is the genome? The genome size will greatly influence the amount of data that must be ordered and analyzed. To assemble a genome, securing a certain number/amount of sequences/depth/coverage (called reads) is the first step before proceeding with ordering sequence data. To get an idea of the size and complexity of a genome, publicly available databases for approximate genome sizes are accessible for fungi (http://www.zbi.ee/fungal-genomesize), animals (http://www.genomesize.com), and plants (http://data.kew.org/cvalues). Selecting a closely related species is a practical option if the information on a target species is unavailable from a public database. Alternatively, the two widely used flow cytometry and *k*-mer frequency distribution methods could provide reliable genome size estimates to predict repeat content and heterozygosity rates. Flow cytometry is a fast, easy, and accurate system of simultaneous multiparametric analysis for nuclear DNA content including a ploidy level that isolates nuclei stained with a fluorescent dye [50,51]. *K*-mer frequency distribution, a pseudo-normal/Poisson distribution around the mean coverage in the histogram of *k*-mer counts, is a powerful and straightforward approach to use raw Illumina DNA shotgun reads to infer genome size, data preprocessing for de Bruijn graph assembly methods (tune runtime parameters for analysis tools), repeat detection, sequencing coverage estimation, measuring sequencing error rates, and heterozygosity [52,53]. It is highly recommended to use both flow cytometry and *k*-mer methods—the gold standard for genome size measures when designing genomic sequencing projects—because no single sequence-based method performs well for all species, and they all tend to underestimate genome sizes [54]. Is it a diploid, polyploid, or highly heterozygous hybrid species? If possible, it is better to use a single individual and sequence a haploid, highly inbred diploid organism [20,23,55], or isogenic line [56] because this will essentially minimize potential heterozygosity problems for genome assembly. While most genome assemblers are haploid mode (some diploid-aware mode) to collapse allelic differences into one consensus sequence, using complex polyploid or less inbred diploid genomes can greatly increase the number of present alleles, which will likely result in a more fragmented assembly or create uncertainties about the contigs' homology [14,49]. If so, polyploid and highly repetitive genomes may require 50% to 100% more sequence data than their diploid counterparts [14].

Is there high/low GC content in a genomic region? Extremely low or high GC content in a genomic region is particularly known to cause problems for second-generation sequencing (SGS) technologies (also called short-read sequencing: mainly refer to Illumina sequencing), resulting in low or no coverage in those regions [57]. While this can be compensated for by increasing the coverage, we would recommend using third-generation sequencing (TGS) technologies (PacBio and ONT) that do not exhibit this bias [14,49].

How many repetitive sequences (or transposable elements) will likely be present in the genome? The amount and distribution of repetitive sequences, potentially occurring at different locations in the genome, can hugely influence genome assembly results, simply because reads from these different repeats are very similar and the assemblers' algorithms cannot distinguish them effectively. This may eventually lead to misassembly and misannotation. This is particularly true for SGS reads and assemblies, and a high repeat content will often lead to a fragmented assembly because the assemblers cannot effectively determine the correct assembly of these regions and simply stop extending the contigs at the border of the repeats [58]. To resolve the assembly of repeats (or if the subject genome has a high repeat content), using TGS reads that are sufficiently long to include the unique sequences flanking the repeats is an effective strategy [14,49]. Thus, understanding the target genome and generating sufficient sequence data/read coverage is a crucial starting point in a genome assembly and annotation project.

Step 3: Design the best experimental workflow

To meet the experimental goals and answer various biological questions, each application must come with different experimental designs. Above all, the development of high-quality chromosomally assigned reference genomes constitutes a key feature for understanding a species' genome architecture and is critical for the discovery of the genetic blueprints for biologically significant traits. Once the reference genome has been completed, follow-up post-genome studies can be substantially completed with high accuracy.

While NGS is a useful tool for determining DNA sequences, certain parameters need to be considered prior to running an NGS experiment, such as quality control, SGS versus TGS, read length, read quality/error rate, number of reads, genome read coverage/depth, library preparation, and downstream applications. Recent papers have provided useful recommendations and strategies to ensure the success of NGS experiments by selecting the correct products/technologies and methods for the project [14,59–61]. If money is no obstacle, using TGS data (PacBio and ONT) and Hi-C data is recommended [14], which are also widely accepted approaches for reaching a chromosome-level genome assembly (Table 1) for aquaculture or any other species. While a hybrid approach using Illumina/10x Genomics Chromium (10xGC) and Hi-C data has been proposed as a cost-effective method, this approach's contiguity could be lower than that of the combination of TGS data and Hi-C data [14].

Another important point to consider is whether genome assembly should be de novo or reference guided/assisted (Table 2). De novo assembly is the most widely adopted, but when complete genomes of closely related species are available, reference-guided/assisted genome assembly could be an attractive option because of its lower requirements for coverage data and computational memory [14]. However, early works have warned against its applications in genome assembly because the resultant assemblies may contain biases toward errors and chromosomal rearrangements in the existing reference genome [62–64]. No matter which assembly approaches and technologies are taken, genome assembly's purpose is to construct a consensus haploid or haploid-phased chromosome-level assembly. Most extensively used genome assemblers typically collapse the 2 sequences into 1 haploid consensus sequence and thus fail to capture the diploid nature of target organisms. While this has been a key challenge in the bioinformatics and biology community, recent works have demonstrated the effectiveness of generating accurate and complete haplotype-resolved assemblies for diploid and polyploid species (Table 2). While we have provided a brief summary of commonly used tools (Table 2), the comprehensive program list focused on TGS reads can be accessed at LRS-DB (https://long-read-tools.org). Thus, selecting the appropriate tools and pipelines is important to achieve accurate chromosome-scale assemblies in a timely manner by leveraging speed and sensitivity in the contiguity and quality of genome assemblies.

Step 4: Choose the best sequencing platforms and library preparations

To sequence an organism's entire genome (WGS), it must be prepared into a sample library from high-quality genomic DNA. A library is a collection of randomly sized DNA fragments that represent the sample input; its size can vary depending on the choice of sequencing technology. Sample library preparation for WGS is dependent on two considerations: (1) the genome size of the target sample organism; and (2) the amount of sample available to be sequenced. Given the vast range of library preparation products, we can only provide general suggestions for library preparations. For more platform-specific library preparation and sequencing guides, refer to the vendor's products and/or services page. The recommended procedure is to select the best and most cost-effective library preparation and sequencing technology after considering the given research goal and budget.

Table 2. Commonly used tools and programs for genome assembly.

Name	Official link	Main feature
De novo genome assemblers for TGS reads		
Falcon/HGAP	https://pb-falcon.readthedocs.io/en/latest/#	Diploid-aware mode including trim, correction, and consensus for PacBio reads
CANU	https://github.com/marbl/canu	A fork of the Celera Assembler including trim, correction, and consensus for TGS reads
SMARTdenovo	https://github.com/ruanjue/smartdenovo	De novo assembler including all-vs.-all raw read alignments without an error correction stage for TGS reads
MECAT	https://github.com/xiaochuanle/MECAT	Ultrafast mapping, error correction, and de novo assembly tools for single-molecule sequencing reads
Flye	https://github.com/fenderglass/Flye	A repeat graph mode including trim, correction, and consensus with polishing for TGS reads
Shasta	https://github.com/chanzuckerberg/shasta	A run-length representation of ONT reads
De novo genome assemblers for SGS reads		
ABySS2	https://github.com/bcgsc/abyss	An assembler intended for SGS PE and linked-reads
AllPath-LG	http://software.broadinstitute.org/allpaths-lg/blog/	Uses a unipath graph from the *k*-mer paths to collapse repeats
MEGAHIT	https://github.com/voutcn/megahit	An ultrafast and memory-efficient assembler for SGS reads
SOAPdenovo	http://soap.genomics.org.cn	De Bruijn graph assembler with an error correction stage
De novo genome assemblers for hybrid reads		
MaSuRCA	https://github.com/alekseyzimin/masurca	An assembler combining the benefits of the de Bruijn and Overlap-Layout-Consensus assembly approaches for SGS and TGS reads
Reference-guided/assistance assemblers		
Ragout	https://github.com/alekseyzimin/masurca	Chromosome-level scaffolding
RaGOO	https://github.com/malonge/RaGOO	Pseudochromosome construction
RGAAT	https://github.com/wushyer/RGAAT_v2	Genome assembly and annotation
Haplotype/phase assemblers		
Falcon-Unzip	https://pb-falcon.readthedocs.io/en/latest/index.html	PacBio reads
Falcon-Phase	https://github.com/phasegenomics/FALCON-Phase	PacBio reads
Triobinning	https://github.com/skoren/triobinningScripts	ONT reads
Platanus-allee	http://platanus.bio.titech.ac.jp/platanus2	SGS and TGS reads
WhatsHap	https://bitbucket.org/whatshap/whatshap/src/master/	SGS and TGS reads
IntegratedPhasing	https://github.com/vibansal/IntegratedPhasing	SGS and TGS reads
HaploConduct	https://github.com/HaploConduct/HaploConduct	SGS and TGS reads
HaplotypeAssembler	https://github.com/ComputationalGenomics/HaplotypeAssembler	SGS and TGS reads

ONT, Oxford Nanopore Technology; PacBio, Pacific Bioscience; PE, paired-end; SGS, second-generation sequencing; TGS, third-generation sequencing.

https://doi.org/10.1371/journal.pcbi.1008325.t002

The rapid adoption of WGS has been facilitated by the development of SGS and TGS technologies, which have dramatically reduced sequencing costs and simplified genome assembly. It is possible to select short (Illumina, 454, SOLiD, and Ion Torrent), long (ONT and PacBio), or a combination (hybrid) read. Comprehensive guidelines (including pros and cons) for selecting the correct sequencing technology have been extensively described in previous works [14,59,61,65]. Briefly, while SGS technologies can produce high-throughput, fast, cheap, and highly accurate reads of lengths in the range 75 to 700 bp, they show limited ability to resolve complex regions with repetitive or heterozygous sequences, which results in incomplete or heavily fragmented genome assemblies. According to Illumina, widely used SGS technology—the TruSeq PCR-free Library Preparation Kit—is ideal for any size of genome with a large sample input if there is 2 μg of genomic DNA available. However, the Nextera DNA Library Prep

Kit (Illumina) is perfect for large and complex genomes with a small sample input. Meanwhile, the TruSeq Nano DNA Library Prep Kit (Illumina) is ideal for any size genome with a small sample input if there is only 200 ng of genomic DNA available. However, the Nextera DNA XT DNA Library Preparation Kit (Illumina) is perfect for small genomes, plasmids, and amplicons. Additional Illumina library preparation methods and sequencing platforms for high throughput have been extensively reviewed [66,67].

Meanwhile, TGS technologies can produce long single-molecule reads (averaging >30 kb) with complete contiguity, facilitating assembly. However, long-read technologies suffer from both high costs per base and high error rates. To overcome this disadvantage, the PacBio RS II or SEQUEL system (Pacific Biosciences, Menlo Park, California, USA) has been released that could generate 10 to 15 times more data than the original SEQUEL system with even more accurate long reads (HiFi reads could be ABI Sanger quality up to 40 kb). According to PacBio, the SMRTbell Template Prep Kit (Pacific Biosciences) with 20 to 40 kb template preparation using BluePippin Size Selection is recommended for WGS [14,68]. For ONT, a combination of ligation sequencing, PCR sequencing, and rapid sequencing has been optimized for WGS [60,69]. In particular, the Rapid Sequencing Kit (SQK-RAD004) could produce even higher read lengths and some reads could be >2 Mb [70].

Combining data from both SGS and TGS in a "hybrid approach/assembly" can compensate for the downsides of both approaches and is a cost-effective method because SGS data can correct errors in TGS reads [33,71–75]. Alternatively, the development of an advanced "hybrid" approach, such as incorporating 10xGC data or medium-size single-molecule DNA fragment selection and tagging before short-read sequencing, could be a practical strategy to increase the continuity and accuracy of long reads [14]. While recent studies have highlighted the efficacy and cost-effectiveness of 10xGC linked-reads in diploid aquatic species' genomes [76–79], the utility of this technology for complex and/or polyploid aquatic species is still being investigated. According to 10xGC, the Chromium Genome Reagent Kit is ideal.

Regardless of the sequencing technology and approach (SGS, TGS, or hybrid), incomplete and/or unfinished assemblies can still occur (e.g., those with gaps and fragments). Thus, additional techniques such as optical mapping (BioNano, San Diego, California, USA) and chromatin association (Hi-C) are highly recommended to facilitate contig joining and genome assembly completion [46,80–83]. Use of the Hi-C method over BioNano has been observed in aquaculture species (Table 1). The most widely used kit is the Proximo Hi-C Kit provided by Phase Genomics (https://www.phasegenomics.com/hi-c-kits).

Step 5: Select the best possible DNA source and DNA extraction method

The extraction of high-quality DNA is the most important aspect of a successful genome project. Given the potential breadth of aquaculture species, each with their own peculiarities, extracted high-molecular-weight DNA should be free of contaminants either from the subjected material itself or from the DNA extraction procedure (e.g., polysaccharides, proteoglycans, proteins, secondary metabolites, polyphenols/polyphenolics, humic acids, carbohydrates, and pigments). While recent publications and commercial kits have provided valuable guidance [84–86], DNA extraction methodologies can be explored and adapted along the lines provided by the literature. In general, the minimum DNA input is required for Illumina and 10xGC > 3 ng, PacBio > 20 μg, ONT > 1 μg, BioNano > 200 ng, and Dovetail > 5 μg [14]. Depending on the project budget and sequencing platform accessibility, SGS and/or TGS technologies can be considered; we recommend using TGS that can deliver DNA of average size >25 kb. Certain species (e.g., mollusks containing high levels of polysaccharide) warrant more careful planning than others. A modified low-salt cetyltrimethylammonium bromide

extraction protocol has produced excellent quality DNA of high molecular weight that is free from contaminants and shearing [87]. Other important considerations are the heterozygosity rate, amplification, and presence of other tissues/organisms [14,49]. The heterozygosity rate can be reduced using a single individual for extraction. However, certain organisms require a pool of individuals to retrieve a sufficient amount of DNA, which will increase the genetic variability and lead to a more fragmented assembly. Attractive strategies include generating an inbred line of individuals for low-heterozygosity pooled sequencing and/or sequencing of haploid tissues as the foundation for filtering out paralogous sequence variants. These have been successful for cost-effective WGS and for optimizing the precision of allele and haplotype frequency estimates in aquaculture breeding [19,20,24,42,55]. When few cells are available, the genomic DNA must be amplified before sequencing, but this can often result in uneven coverage due to artificial effects (chimeric and/or fused unrelated sequences). The introduction of unwanted/unrelated organisms (e.g., contaminants and/or symbionts) and/or tissues (e.g., mitochondria and/or chloroplasts) should be minimized at the extraction and library preparation stages. This requires using tissue with a higher ratio of nuclear over organelle DNA because this can lead to higher coverage of the nuclear genome in the sequences. Whichever approach is adopted, there will be a need to refine the method to achieve several important quality metrics for genome sequencing.

Care should be taken for quality parameters (e.g., the chemical purity and structural integrity of DNA) and two recent works have made the recommendations outlined below for long-read technologies [14,49]. Generally, the measurement/quantification of purified DNA should be performed using both spectrophotometric and fluorescence-based methods (e.g., qubit). Samples with optical density (OD_{260}:OD_{280}) ratios of 1.8 to 2.0 are usually free of protein contamination. DNA concentrations at a 1:1 ratio (determined by spectrophotometry and fluorimetry, respectively) are very good indicators of whether they will be sequenced efficiently. To determine the integrity of DNA samples, contour-clamped homogeneous electric field or pulsed-field gel electrophoresis is appropriate when used with TapeStation or Fragment Analyzer (Agilent Technologies, Santa Clara, California, USA). Analyzing isolated DNA in this manner also facilitates decisions regarding shearing DNA to attain an optimal size range for sequencing. Thus, it is always worth investing time in getting high-quality DNA that will result in high-quality data and assembly to save time and money.

Step 6: Check the computational resources and requirements

Installing open-source tools in one's computational environment is not always either straightforward or trivial. It generally poses three potential problems: (1) the prerequisites of the tools created by diverse developers employing diverse programming frameworks differ; (2) the installation of various software items in one environment can lead to hard-to-resolve software dependency conflicts; and (3) upon successful installation, maintaining the environment and ensuring that all tools (including changes and updates) are working as expected remain difficult. Therefore, managing the data analysis environment becomes increasingly complex when a project requires many tools for genomic data analysis. While addressing the importance of the appropriate data and computing infrastructure to genome projects is difficult, the two following options (see Step 7: maximizing in-house workers or collaboration and outsourcing from the service provider) can be considered.

Access to high-performance computing or cloud-based computing systems is crucial for genome projects that require a large number of computing resources. As a general guide, the successful assembly of a moderately sized diploid genome (approximately 1 Gb) using software pipelines (Tables 1 and 2) requires a minimum computing resource of 96 physical central

processing unit (CPU) cores, 1 TB of high-performance random-access memory (RAM), 3 TB of local storage, and 10 TB of shared storage [14]. However, the guide is scalable based on the amount of data, genome size, heterozygosity rate, and ploidy. Please note that runtimes, memory requirements, number of CPUs, and computational costs will increase geometrically because genome assembly is an all-by-all comparison. However, hard drive space to store raw and/or intermediate data (e.g., storage space) will increase linearly as the total amount/depth of coverage required does not dramatically change as genomes increase in size. In addition, the recommendations stated here will likely apply to larger and more complex genomes (e.g., crustaceans with numerous chromosomes) but at a slower rate and with higher computing resources and costs (obtaining more computing resources will increase costs). If participants' or collaborators' institutions are equipped with large in-house high-performance computing resources, they will likely have more direct access and practical assistance in their genome project. Otherwise, cloud-based computing is a potential solution that has been widely emphasized in previous works including easy-to-follow steps [88–90]. While cloud computing provides flexibility, competitive pricing, and continually updated hardware and software, it still requires assistance from information technology (IT) specialists to set up suitable cloud-based software. Thus, users should consider all possible options (including their research budget) to achieve the best outcome.

Step 7: Choose the best computational design and pipeline

Optimizing a computational design and securing sufficient computer resources are essential steps to succeed in a genome assembly and annotation project. In addition, computational proficiency and literacy have become vital skills for biologists to design and interpret big data analyses and multi-omics studies [48]. Given the vast range of computational tools and requirements (different resource demands between assembly and annotation for each species), general suggestions are provided on the computational aspect. However, when establishing the best and most cost-effective computational design and requirement, it is important to consider three options: (1) maximizing in-house workers or collaboration; (2) outsourcing from a service provider; and (3) simulating data with different settings. Ultimately, the most suitable and practical approach in methodological computational biology research is recommended because there is no perfect computational design for genome assembly and annotation.

Before embarking on any actual data analyses, the overall goals should first be defined by understanding in-house workers and facilities because computational design requires extensive learning of computer and biology knowledge, which is a great challenge for most wet lab researchers/groups. If in-house workers and computer facilities are not ready to deliver successful outcomes, cross-disciplinary collaborations (computer science, data science, bioinformatics, and biology) could present great solutions. Initiating and successfully maintaining cross-disciplinary collaborations can be challenging but are highly rewarding because the combination of methods, data, and interdisciplinary expertise can achieve more than the sum of the individual parts alone [91].

Alternatively, work can be outsourced to a service provider. Outsourcing has the following benefits: (1) no need to hire more employees for computational design and analysis, which will reduce labor costs; and (2) there are more talents available at well-equipped companies that are very specialized in specific research fields. However, outsourcing also has the following disadvantages: (1) a lack of control as a contractor; (2) limited methods of communication (e.g., phone, e-mail, or online chat); and (3) the potential danger of poor quality work due to the inability to optimize pipelines (e.g., parameters) and outcomes.

No matter which approach is taken, the essential part is to have firsthand experience to select proper computational design and pipeline and to accurately interpret analyzed genome

data. Due to its extensive range of analytical tools and application areas, employing an effective simulator (from the quality of raw reads to assembly evaluation) has become an essential step for benchmarking genomic and bioinformatics analyses [92–94]. In simulations, considering a (very) large number of datasets is generally not a problem, except when the analysis of each dataset is hugely computationally expensive (e.g., in the genome assembly stage). In practice, one should generate and analyze as many datasets as computationally feasible before embracing real empirical studies, particularly before undertaking real assemblies. In large genome assembly, simulating assemblies of down-sampled real data (e.g., 30× coverage/depth of genome) would be very useful for selecting the best pipeline and parameters without requiring too much computational time or cost. Ultimately, a simulation's practical relevance depends on the similarity between the considered simulation settings and the real datasets in the area of application. The new method may be assessed in different ways depending on the context (e.g., by conducting simulations, applying the method to several real datasets, applying flexible parameter settings, and checking the underlying assumptions in practical examples). Therefore, simulations should not be limited to artificial datasets that correspond exactly to the assumptions underlying the new method as this would favor the new method [61,95–98].

Step 8: Assemble the genome

Regardless of which pathway/strategy is chosen, the TGS approach is recommended over the SGS or hybrid approaches. In general, using multiple programs at each stage to predict the best assembly and annotation (Table 2) is also recommended because each approach and tool has limitations based on the problems inherent in the different algorithms and assumptions used. If the abovementioned steps (Steps 1–7) are met, the recommended flowchart and/or guideline for genome assembly, annotation, maintenance, and community effort would be as shown in Fig 1, which could be broadly applicable to any species. The rationale of each computational design, workflow, and decision tree is well described in Jung and colleagues [14],

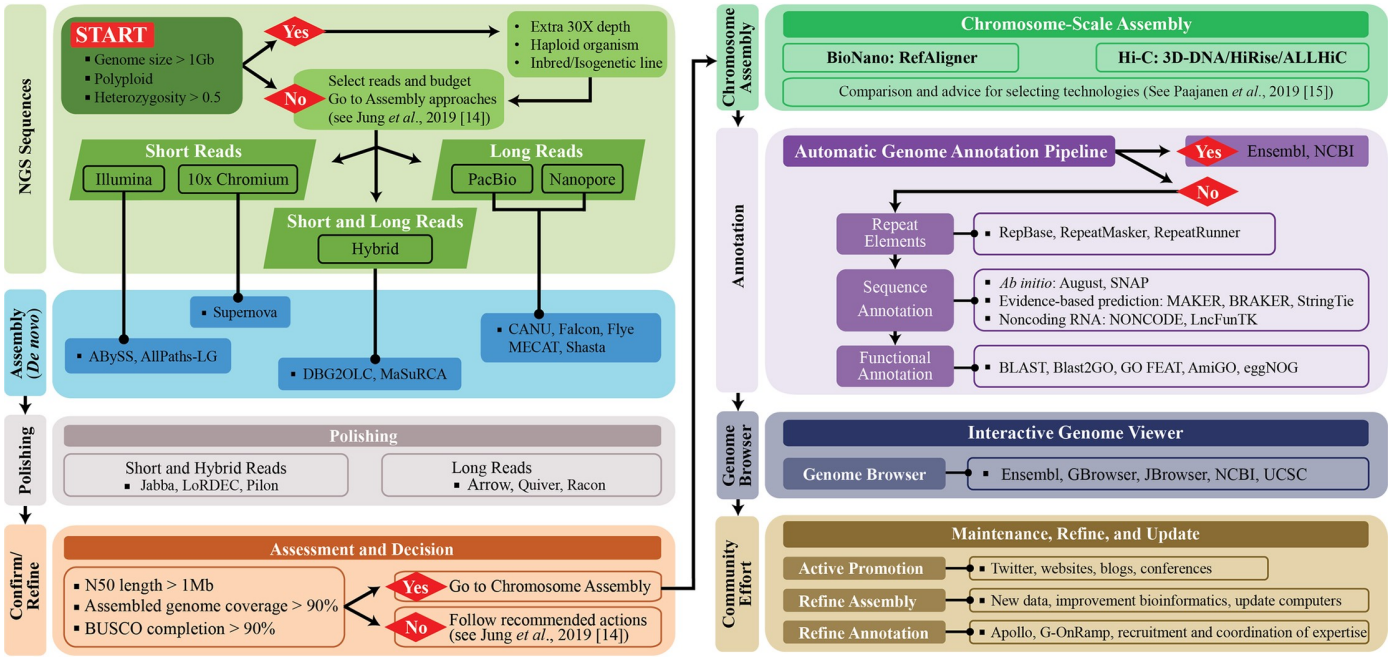


Fig 1. Recommended flowchart for genome assembly and annotation. NGS, next-generation sequencing.

https://doi.org/10.1371/journal.pcbi.1008325.g001

including the background information for each of their steps and the spectrum of available analytical options. Following the workflow and decision tree described by Jung and colleagues, the recommended tools herein are the TGS pipeline: PacBio/ONT read sequencing (remove all contaminated DNA; plastids/bacterial contamination) → read quality assessment, evaluation, and filtering → assembly → error correction and polishing using SGS reads → assessment → chromosome-level assembly using BioNano and Hi-C data. Several recent assemblies adopted from this pipeline (or similar) have shown notable improvements in the assembly of intergenic spaces and centromeres [33,72]. A potential assembly outcome from the new SEQUEL II (HiFi) reads would be even more promising (see Step 4) compared to its early version SEQUEL. In the SGS pipeline, if the target is a diploid organism, starting from 10xGC read sequencing over Illumina reads is ideal. Based on the results of the hybrid-based assemblies, the recommended pipeline starts from PacBio/ONT, and 10xGC read sequencing greatly helps build a highly accurate contiguous genome [78]. However, all assembly approaches/designs derived only from sequence reads will still contain misassemblies (inversions and translocations), these are mainly caused by the inability of both sequencing and assembly pipelines to cope with long tracts of repeat sequences or high levels of heterozygosity and polyploidization. Thus, using BioNano and Hi-C data is highly recommended for reaching chromosome-level assembly because these two methodologies/technologies can improve the assembly quality by validating the integrity of the initial assembly, correcting misorientations, and ordering the scaffolds.

Step 9: Check the assembly quality before annotation

In the shotgun sequencing era, assembling a new genome mostly relies on computational algorithms and experimental designs (see Steps 6 and 7). The performance of such algorithms and designs, read lengths, insertion size of sequencing libraries, read accuracy, and genome complexity determines the accuracy and continuity of the genome assembly. Therefore, while estimating assembly quality is an unpredictable and challenging task that requires several statistical and biological validations, it remains an important step for a high-quality genome. Typically, the quality assessment for draft assemblies is carried out via statistical measurements and alignment to a reference genome (if available) [99]. These include overall assembly size (determining the match to the estimated genome size), measures of assembly contiguity (N50, NG50, NA50, or NGA50; the number of contigs; contig length; and contig mean length), assembly likelihood scores (calculated by aligning reads against each candidate assembly), and the completeness of the genome assembly (Benchmarking Universal Single-Copy Orthologs [BUSCO] scores and/or RNA-seq mapping) [100,101]. In computational biology, N50 is a widely used metric for assessing an assembly's contiguity, which is defined by the length of the shortest contig for which longer and equal-length contigs cover at least 50% of the assembly. NG50 resembles N50 except for the metric, which relates to the genome size rather than the assembly size. NA50 and NGA50 are analogous to N50 and NG50 where the contigs are replaced by blocks aligned to the reference [99]. Thankfully, recent bioinformatics tools offer an automated pipeline to compute and evaluate the new genome quickly and accurately in a practical setting [44,102,103].

Additional strong indicators of quality include agreement with data on quantitative trait loci, expressed sequence tags (ESTs), fluorescent in situ hybridization experiments employing bacterial artificial chromosome clones, and the genome assembly's contiguity with a chromosome-level genetic map. If the initial assembly attempt is unsatisfactory, three specific areas (contiguity, accuracy, and completeness) should be considered to determine the best path forward to improve the new assembly's quality [14]. Generally, the best way to address high contig numbers with low average size is to acquire and incorporate more TGS or 10xGC (see Steps

3 and 4: hybrid assembly approaches) reads. When attempting to increase assembly quality, adding more and longer TGS reads tends to be more helpful for bridging existing contigs by increasing the size of the average contig; then, subsequently adding further BioNano and Hi-C data improves read accuracy and assemblies' overall contiguity. Unfortunately, additional BioNano and Hi-C data without TGS reads are unlikely to help increase the assembly quality because the data are usually ineffective at assisting hybrid assemblers span gaps between existing contigs [14]. To obtain a complete genome, applying LR_Gapcloser, a fast and memory-efficient approach using long reads, would be an excellent choice to close gaps and improve the contiguity of genome assemblies [104].

Step 10: Genome annotation

Unlike advanced and revolutionized genome sequencing and assembly, getting genome annotation correct remains a challenge. Annotation is the process of identifying and describing regions of biological interest within a genome (both functionally and structurally). While there are various online annotation servers (Table 3), the intended use of the curated data needs to be clearly defined after considering the two options addressed in Step 7 (maximizing in-house workers/collaboration and outsourcing) because the gene-finding problem in eukaryotes is far more difficult than that in prokaryotes such as bacteria. This procedure requires advanced bioinformatics skills, pipelines, and computing resources and consists of three main steps: (1) identifying noncoding regions; (2) identifying coding regions (called gene prediction); and (3) attaching the biological information of these elements.

Recent works have described genome annotations well [13,105–109]. However, it is highly recommended that beginners select automatic or semiautomatic annotation methods (including the workflow and guideline in Fig 1) because manual annotation can be very time- and labor-intensive and expensive. Note that while automatic procedures help accelerate the annotation process, they decrease the confidence and reliability of the outcomes because results from different servers and/or databases are often dissimilar [106,110,111]. Furthermore, automatic annotation algorithms, frequently based on orthologs from distantly related model organisms, cannot yet correctly identify all genes within a genome and manual annotation is often necessary to obtain accurate gene models and gene sets [106,110,111]. Thus, a scheme to obtain consensus annotations by integrating different results, a semiautomatic method, is in demand because this could balance automatic and manual approaches, which would increase the reliability of the annotation while accelerating the process [106,110,111]. In general, the identification of noncoding regions includes small and long sequences including repetitive and transposable elements (Fig 1 and Table 3). Despite an explosion of interest in noncoding data and the massive volume of scientific data, selecting the best strategy to annotate and characterize noncoding RNAs is a daunting task because of the strengths and weaknesses of each computational and empirical approach [112]. After screening noncoding regions (e.g., repeat masking and transposable elements), elements of the gene structure (e.g., introns, exons, coding sequences [CDSs], and start and end coordinates) can be predicted for coding regions.

Both ab initio and evidence-based prediction approaches are widely used as each approach has pros and cons. While Augustus and SNAP are the most popular tools for ab initio prediction, they still necessitate the information of the closely related gene and genome model for screening against the newly sequenced genome. By contrast, evidence-based prediction usually uses results obtained by aligning ESTs, protein sequences, and RNA-seq data (results are even better with full-length Iso-Seq data from PacBio or ONT) to a genome assembly as external evidence. Trained gene predictors (training with Augustus and SNAP to obtain more accurate annotation results is highly recommended) can be used in MAKER, BRAKER, and StringTie

Table 3. Commonly used genome annotation tools and programs.

Name	Official link	Main feature
Online pipeline		
NCBI	https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/	Eukaryotic genome annotation. An automatic pipeline with flexibility and speed. Good for beginners and easy to use.
	https://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards/	Prokaryotic genome annotation. An automatic pipeline with flexibility and speed. Good for beginners.
Ensembl	http://ensemblgenomes.org/info/data/annotation https://asia.ensembl.org/info/genome/genebuild/assembly.html	Genome annotation. An automatic pipeline for importing external data or using predictive algorithms. Good for beginners and easy to use. Annotation and prediction.
GenSAS	https://www.gensas.org	Integrates with JBrowse and Apollo. An automatic platform and pipeline for genome structural and functional annotation. A user-friendly interactive portal that includes visualization and editing. Good for beginners and easy to use.
GO FEAT	http://computationalbiology.ufpa.br/gofeat/	Genome and transcriptome. A rapid automatic platform for functional annotation and enrichment. A user-friendly portal that can export results in different output formats. Good for beginners and easy to use.
Blast2GO	https://www.blast2go.com	Functional annotation. An automatic platform as a standalone application that has high throughput and is interactive. A user-friendly program with easy start-up and low maintenance. Good for beginners, but the pro version requires a commercial license.
AmiGO	http://amigo.geneontology.org/amigo	GO and GO enrichment analysis. A user-friendly web-based platform. Requires some configuration of public databases with Perl, JavaScript, and Linux for the standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
eggNOG	http://eggnogdb.embl.de/#/app/home	Database of orthologous groups and functional annotation. An automatic platform and pipeline for any genome that scales with speed and flexibility (15 and 2.5 times faster than BLAST and InterProScan, respectively). Requires some configuration of public databases with various computer languages for a standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
KAAS	https://www.genome.jp/tools/kaas/	Ortholog assignment and pathway mapping. An automatic platform but has a limited number of query sequences. A good web resource for beginners, but local installation requires bioinformatics support.
Augustus	http://bioinf.uni-greifswald.de/augustus/	Gene/genome structure and annotation using ab initio and transcript-based prediction. An automatic platform and pipeline for eukaryotic genomes. Requires some configuration of public databases with various computer languages and dependencies for a standalone application. A good web resource for beginners, but local installation requires bioinformatics support.
GAAP	http://GAAP.hallym.ac.kr	A semiautomated genome assembly and annotation pipeline.
Command line interface		
BRAKER	https://github.com/Gaius-Augustus/BRAKER	Gene/genome structure and annotation using a combination of GeneMark-ET, Augustus, and RNA-seq evidence. A fully automated training platform for novel eukaryotic genomes. Requires 2 input files: an RNA-seq alignment file in BAM format and a corresponding genome file in fasta format. Good for intermediate and advanced users due to the requirement of several semi-unsupervised pipelines and dependencies in local installation.
MAKER	https://www.yandell-lab.org/software/maker.html	Gene/genome structure and annotation pipeline. An easy-to-use semiautomatic pipeline for the de novo annotation of newly sequenced genomes for updating existing annotations to reflect new evidence or just to combine annotations, evidence, and quality control statistics for use with other GMOD programs such as G/JBrowse, Chado, and Apollo. Good for intermediate and advanced users due to the requirement of several semi-unsupervised pipelines and dependencies in local installation.
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/	Transcriptome assembly and differential expression analysis of RNA-seq. A semiautomatic pipeline that includes TopHat (read mapping) and CummeRbund (visualization and exploration). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
StringTie	https://ccb.jhu.edu/software/stringtie/	A fast and highly efficient assembler of RNA-seq alignment that allows users to quantitate full-length transcripts representing multiple splice variants for each gene locus. A semiautomatic pipeline using a BAM alignment input file with RNA-seq read mappings (produced and converted by TopHat, HISAT2, and Samtools). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
GLEAN	https://sourceforge.net/projects/glean-gene/	An unsupervised learning system for gene structure prediction. A semiautomatic pipeline without prior training. Lacks proper documentation and resources to run programs. Might be good for advanced users due to the requirement of several pipelines and dependencies in local installation.

(*Continued*)

Table 3. (Continued)

Name	Official link	Main feature
BLAST	https://blast.ncbi.nlm.nih.gov	A specialized algorithm to find regions of local similarity between sequences. A semiautomatic pipeline for understanding biological sequences. A good web resource for beginners, but local installation requires bioinformatics support.
Modeler	https://evidencemodeler.github.io	Software combining ab initio gene predictions and protein/transcript evidence into weighted consensus gene structures. A semiautomatic pipeline with a flexible and intuitive framework for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
GSNAP	http://research-pub.gene.com/gmap	A genomic mapping and alignment program for mRNA and ESTs. A semiautomatic pipeline for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
SNAP	https://github.com/KorfLab/SNAP	Semi-HMM-based nucleic acid parser gene prediction tool. A semiautomatic pipeline for gene structure annotation. Good for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
TopHat	https://ccb.jhu.edu/software/tophat/index.shtml	A fast splice junction mapper for RNA-seq. A semiautomatic pipeline that includes Bowtie and HISAT2 (read aligner). Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
PASA	https://github.com/PASApipeline/PASApipeline/wiki	Program for assembling spliced alignments for genome annotation and gene structures. A semiautomatic pipeline for gene structure annotation but useful for genome-guided and de novo RNA-seq assemblies to generate a comprehensive transcript database. Good for intermediate and advanced users due to the requirement of several pipelines and dependencies in local installation.
Evigan	http://www.seas.upenn.edu/~strctlrn/evigan/evigan.html	Predicts genes by integrating multiple evidence sources. An automated annotation program that employs a Dynamic Bayesian Network. Model parameters are estimated by the Expectation–Maximization algorithm, thus eliminating the need to curate training data. Good for intermediate users due to the local installation requirement.
Noncoding RNAs		
Ensembl	https://asia.ensembl.org/info/genome/genebuild/ncrna.html	Automatic annotation of noncoding genes but requires registration. A good web resource for beginners.
LncFunTK	http://sunlab.cpy.cuhk.edu.hk/lncfuntk/	Functional annotation of long noncoding RNAs. An easy-to-use automatic pipeline for newly assembled genomes but requires several input files such as expression profiles (GTF format), TF binding profiles (BED format), and miRNA-binding profiles. This is a good web resource for beginners but might be better for intermediate and advanced users due to the requirement of several input files, pipelines, configurations, and dependencies in local installation.
NONCODE	http://www.noncode.org	Database for noncoding RNAs except tRNAs and rRNAs. An automatic pipeline including 6 steps, format normalization (BED or GTF), combination, filtering protein-coding RNA, information retrieval, advanced annotation, and web presentation. This has a good user-friendly web interface for beginners, but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
deebBase	http://rna.sysu.edu.cn/deepBase/	Small RNAs, lncRNAs, and circular RNAs
lncRNAdb	https://rnacentral.org/expert-database/lncrnadb	A database that provides comprehensive annotations of eukaryotic long noncoding RNAs. An easy-to-use open public resource. An automatic pipeline for single sequences and a semiautomatic pipeline for multiple sequences with bioinformatic scripts. This has a good user-friendly web interface for beginners but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.
Repeat element		
RepeatMasker	http://repeatmasker.org	A program to screen for interspersed repeats and low-complexity DNA sequences. A fast and sensitive semiautomatic pipeline for assembled genomes. Good for intermediate and advanced users due to the requirement of several databases, pipelines, and dependencies in local installation.
RepeatRunner	http://www.yandell-lab.org/software/repeatrunner.html	A CGL-based program that integrates RepeatMasker with blastx to identify repetitive elements. A semiautomatic pipeline for assembled genomes. Good for intermediate and advanced users due to the requirement of several databases, configurations, pipelines, and dependencies in local installation.
RepBase	http://www.girinst.org/repbase/update/index.html	A database of prototypic sequences representing repetitive DNA from different eukaryotic species. A semiautomatic pipeline for genome sequencing projects. This has a good user-friendly web interface for beginners but it might be better for intermediate and advanced users due to the requirement of several pipelines, configurations, and dependencies in local installation.

BAM, binary alignment map; BED, browser extensible data; ESTs, expressed sequence tags; GO, gene ontology; GTF, gene transfer format; HMM, hidden Markov model; RNA-seq, RNA sequencing; TF, transcription factor.

https://doi.org/10.1371/journal.pcbi.1008325.t003

(Fig 1 and Table 3). When extrinsic evidence from RNA-seq and protein homology information is available, any program/pipeline could be useful for the de novo annotation of novel genomes. In particular, if any RNA-seq data and a genome sequence are available, starting from MAKER and BRAKER over StringTie would be a better choice for a first-time user because MAKER and BRAKER include ab initio prediction (e.g., Augustus training) unlike StringTie (evidence-based prediction only). However, MAKER could be a better choice for updating existing annotations to reflect new evidence. If various gene prediction methods and tools are used to derive the gene structure from a genome, combining these results to obtain the single consensus gene structure via Evidence Modeler, GLEAN, Evigan, or GAAP is essential (Table 3). In particular, BRAKER, StringTie, PASA, and GAAP can update any gene structure annotation by correcting exon boundaries and adding untranslated regions and alternatively spliced models based on assembled transcriptomic data. The evolutionary rapid emergence of new genes (which quickly respond to changing selection pressures) could give rise to orphan genes that might share no sequence homology to genes in closely related genomes [113]. Combining the methods and results (especially MAKER, BRAKER and StringTie) could therefore prove effective in increasing the number and accuracy of annotation predictions assigned to orphan and any other young genes.

Subsequently, functional annotation—the process of attaching biological information to gene or protein sequences—must be performed. This can be carried out through homology search and gene ontology (GO) term mapping. To investigate gene function or predict evolutionary associations, newly assembled sequences should be compared with gene sequences with known functions to find sequences with high homology using BLAST, Cufflinks, TopHat, GSNAP, Blast2GO/OmicsBox (referred to here as Blast2GO), and GAAP (Fig 1 and Table 3). To label more diverse biological information, GO term mapping should be performed, which allows information about gene-related terms and relations between genes to be stored in three categories: biological processes, molecular functions, and cellular components. Mapping is the process of retrieving GO terms associated with hits (mapping sequences) obtained via a previous homology search (mainly BLAST) that are accessible from AmiGO, Blast2GO, GO-FEAT, and eggNOG-Mapper. Starting from Blast2GO would be a practical choice for a complete novice because it has more graphic user interface mode with explanations.

While Fig 1 and Table 3 provide a summary of useful tools with key features, it is highly recommended to be familiar with the regular update of public databases and pipelines. In addition, understanding the performance and capability of various analysis from a detailed comparison and instructions of common features of annotation tools could be a very important factor for a successful genome annotation, structurally [7,111,114–117] and functionally [8,118–123].

Step 11: Build a searchable and sharable output format

Research papers and data products (researchers are usually required to submit raw sequencing data to appropriate repositories such as Sequence Read Archive [SRA]) are key outcomes of the scientific enterprise, including most successful genome projects. In addition, most genomic projects/data potentially have value beyond their initial purpose but only if shared with the scientific community, including refining assembly and annotation (see Step 12). In recent years, genomic studies have involved complex datasets such that biologists have become "big data practitioners" [124] because of improvements in high-throughput DNA sequencing and cost reductions. As a result, genomic studies have become routine procedures, and there is widespread demand for tools that can assist in the deliberative analytical review of genomic information. What happens to the data after such projects end? In general, data or data management plans have become the central currency of science because open access, open data,

and software are critical for advancing science and enabling collaboration across multiple institutions and throughout the world and increasing public awareness [125]. For example, when archiving sequencing data, repositories such as those run by the National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EBI) both provide locations for data archiving and encourage a set of practices related to consistent data formatting and the inclusion of appropriate metadata. However, this is a difficult task for an individual research group due to the wide variety of data formats, dataset sizes, data complexity, data use cases, ethical questions, and data collection/storage/sharing practices [124,126–128]. Despite its importance, major barriers remain to sharing data, software, and research products throughout the scientific community because of the difficulties that interdisciplinary and/or translational researchers face when engaging in collaborative research [124,125,127]. To this end, recent works have provided principles that can be applied in genomic data/database projects, including data sharing and archiving via collaborations [124–128].

The following three fundamental questions on this topic should be considered: (1) Do you want to share your data? (2) Do you have enough in-house expertise and infrastructure to maintain and improve the data, including data storage space? (3) Do you want to form internal and external collaborations to increase research productivity? While each research group has different experiences and criteria in collaborations that included data sharing, engaging with multisite collaborations is highly recommended to overcome more pitfalls, including open-ended questions/concerns on genomic data. In addition, sharing open genomic data can easily facilitate reproducibility and repeatability by reusing the same genomic data.

Step 12: Reach out to the community to refine the assembly and annotation

Dropping whole-genome shotgun sequencing costs and improvements in bioinformatics pipelines and computer capabilities have resulted in the situation where a small lab can undertake genome projects (assembly and annotation), and any organism can become a model species. Ironically, the ease of sequencing and assembly presents another challenge for annotation: contamination of the assembly itself, because errors in assembly can cause errors in the annotation (structural and functional). In addition, it is important to ensure that methods are computationally repeatable and reproducible because there have been numerous reports of instability arising from a mere change of Linux platform, even when using the exact same versions of genomic analysis tools [49]. When including new data, it is also necessary to provide software infrastructure to assist in genomic data updating. Hence, assembled genomes and curated annotations should not and cannot be considered perfect, static, or "final products." Data must be maintained, refreshed, and updated to ensure their reuse and discovery.

Manual and continuous annotation is critical to achieving reliable gene models and elements; however, this process can be daunting and cost prohibitive for small research communities. While some genome consortia choose to manually review and edit sets via time- and resource-intensive meetings that often require substantial expertise, this still provides opportunities for community building, education, and training. In contrast, for small research groups, it has been proposed that involving undergraduates in community genome annotation consortiums can be mutually beneficial for both education and genomic resources [106]. Alternatively, a collaborative approach using web portals such as Apollo, JBrowse, G-OnRamp (Galaxy-based platform), and ORCAE [129–133] could be sufficiently robust and flexible to enable the members of a group to work simultaneously or at different times to improve the biological accuracy of annotation.

Despite any community-based participatory research approaches taken, the recruitment and coordination of researchers are central to any research project due to the requirement of

diverse expertise and collective learning. The ideal way would be to form a national/international collaborative research partnership with diverse organizations [19,134–136]. Alternatively, active promotion via social networks and/or web portal setup could be the most effective way (e.g., Twitter, the Ensemble website, and blogs). Finally, build collective research solidarity by attending conferences would be plausible. There have been previous successful community efforts and involvement in plant (https://nbenth.com/annotator/index, https://solgenomics.net, and https://www.helmholtz-muenchen.de/pgsb) and animal genome projects (http://www.slimsuite.unsw.edu.au/servers/apollo.php, https://bovinegenome.elsiklab.missouri.edu, http://www.gmgi.org/genomics-fish-shellfish, and https://www.sanger.ac.uk/science/data/vertebrate-genomes-sequencing) using the Apollo instance with J Browsers exhibits attractive and effective routes because it is always online, curators can log in whenever they have time, and some minor revisions only require a few seconds (to confirm the gene models). Others require up to 20 minutes to change (UTR boundaries and other structural alterations).

After the initial setup, tasks include maintaining momentum and morale, according to the recommendations described by Pedro and colleagues [137]. Participants bring their own experiences and strengths into this effort. Availability of a training webinar (e.g., https://bit.ly/3gauwn7 and https://bit.ly/36iNQds) would greatly help kick-start the process, alongside a clear set of starting tasks (e.g., a list of genes/families or regions assigned to each curator) and engagement by the community leader. The leader—an enthusiastic champion—can (1) drum up support from their collaborators; (2) fuse community expertise with resources; (3) oversee the project; and (4) act as a liaison between new members wanting to join, the infrastructure provider, and existing annotators. Considering that the collective expertise within a group may be extensive but diverse, it is necessary to standardize the curation for quality control of annotations. To minimize any conflicts that may arise during the annotation process, it is important (1) to have the initial training webinar by laying out clear rules and guidelines; (2) to select a small subset of genes and ask a group of experienced curators to evaluate whether the decisions taken in each case were uniform and sensible; (3) to record webinar training and comments regarding consensus or disagreements for reporting back to the curation team and to edit the tutorial and guidelines; (4) to address this by automated checks and controls (Apollo does not allow this for now or makes it extremely difficult); and (5) to ask multiple reviewers to check each region by reviewing the annotation history in Apollo (labor-intensive method).

Pooling the expertise, resources, and time of active communities could enable a wide range of geographically distance members to participate in a common process, to share and validate the identification of contradictions and the misrepresentation of data on the genomes [137]. After corrections, the datasets (manually verified gene sets) that emerge from these projects can be used to improve the gene sets for closely related genomes and downstream analysis. Dialog and collaboration between community members have an enormous impact. The result of an entire community agreeing on and taking ownership of a single gene set is a major stepping-stone to accelerating the field. Handling the mammoth task of manual gene annotation in the absence of dedicated funding or teams is a great challenge. However, our guidelines could provide a manageable solution for the prospect of this approach becoming commonplace and will continue to engage in community-driven curation efforts.

Advice for new genomic users to select a basic assembly and annotation pipeline

For a complete novice, our recommendation would be as below (not recommended starting from Illumina only short reads assembly).

(1) Pure long-read assembly: PacBio or ONT read sequencing (if combined, PacBio 40X and ONT 25X, or 60X for a single platform) → CANU assembler (alternatively Flye) → BUSCO assessment → Make a decision to add more sequencing data or proceed next step (See Confirm and Refine in Fig 1) → Optional BioNano with RefAligner (still expensive compared to Hi-C data) → Hi-C with 3D-DNA (alternatively HiRise or AllHiC) → Gap-closing with LR_Gapcloser → Arrow with long-read (alternatively Racon) or Pilon polisher with short-read → BUSCO assessment.

(2) Hybrid assembly: 10xGC read with Supernova → PacBio or ONT read with CANU (alternatively MaSuRCA) → The rest are same with "Pure-read assembly" from BUSCO assessment to BUSCO assessment.

(3) Annotation: NCBI or EBI (a web-based automatic pipeline) → If not, proceed a semiautomatic pipeline starting from structural annotation → RepeatMasker → Ab initio Augustus training with MAKER (alternatively BRAKER) → Evidence-based prediction (RNA-seq) with MAKER (alternatively BRAKER) → Noncoding RNA prediction with NONCODE → Functional annotation with Blast2GO (alternatively AmiGO) → Genome Browser.

Conclusions

There are no gold standards for genome assembly and annotation. However, the availability of NGS data (particularly TGS data) and their analytical tools has enabled the sequencing of several high-quality genomes of species of importance in aquaculture in recent years. Beginners and small research groups still face challenges, because genome assembly and annotation are usually complex analytical procedures (or pipelines) requiring interdisciplinary collaborations (from biology to computer science) and hefty costs for refining/maintaining the genome. The recommendations addressed here are broad guidelines that could be considered to avoid common pitfalls throughout the whole-genome assembly and annotation process. However, the comprehensive features (e.g., advantages and disadvantages) of each step and/or technology have not been extensively discussed.

Finally, newly emerging technologies and analytical tools could dramatically improve end-to-end genome assemblies and annotations in the future by replacing the years-long efforts of the past with rapid and low-cost solutions. Meanwhile, emphasis should be placed upon the following: First, define the achievable research aim. Second, avoid the trap of trying to secure a perfect/complete genome assembly and annotation, which could lead to a never-ending project. Third, perform assembly and annotation to gain firsthand experience, including in bioinformatics. Fourth, seek internal and external help and advice from experts. Lastly, be open to sharing genomic data to both increase research productivity and promote public awareness.

Acknowledgments

The authors are grateful to their colleagues, collaborators, and field/technical specialists from each company for their valuable comments.

References

1. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotec. 2020; 18:9–19. https://doi.org/10.1016/j.csbj.2019.11.002 PMID: 31890139
2. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020; 38. https://doi.org/10.1038/s41587-020-0503-6 PMID: 32686750
3. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2020; 20:159.

4. Hatje K, Mühlhausen S, Simm D, Kollmar M. The Protein-Coding Human Genome: Annotating High-Hanging Fruits. BioEssays. 2019; 41:1900066.
5. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic Analysis in the Age of Human Genome Sequencing. Cell. 2019; 177:70–84. https://doi.org/10.1016/j.cell.2019.02.032 PMID: 30901550
6. Chin C-S, Khalak A. Human Genome Assembly in 100 Minutes. bioRxiv. 2019:705616.
7. Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci. 2019; 7:41–64. https://doi.org/10.1146/annurev-animal-020518-115005 PMID: 30379572
8. Bick JT, Zeng S, Robinson MD, Ulbrich SE, Bauersachs S. Mammalian Annotation Database for improved annotation and functional classification of Omics datasets from less well-annotated organisms. Database. 2019; 2019:baz086.
9. Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol. 2019; 17:108. https://doi.org/10.1186/s12915-019-0726-5 PMID: 31884969
10. Giuffra E, Tuggle CK, Consortium F. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu Rev Anim Biosci. 2019; 7:65–88.
11. Rice ES, Green RE. New Approaches for Genome Assembly and Scaffolding. Annu Rev Anim Biosci. 2019; 7:17–40. https://doi.org/10.1146/annurev-animal-020518-115344 PMID: 30485757
12. Etherington GJ, Heavens D, Baker D, Lister A, McNelly R, Garcia G, et al. Sequencing smart: *De novo* sequencing and assembly approaches for non-model mammals. bioRxiv. 2019:723890.
13. Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma B, Faino L. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. Plant Physiol. 2019; 179:38–54. https://doi.org/10.1104/pp.18.00848 PMID: 30401722
14. Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P. Tools and Strategies for Long-Read Sequencing and *De Novo* Assembly of Plant Genomes. Trends Plant Sci. 2019; 24:700–724. https://doi.org/10.1016/j.tplants.2019.05.003 PMID: 31208890
15. Paajanen P, Kettleborough G, Lopez-Girona E, Giolai M, Heavens D, Baker D, et al. A critical comparison of technologies for a plant genome sequencing project. Gigascience. 2019; 8:giy163. https://doi.org/10.1093/gigascience/giy163 PMID: 30624602
16. Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, Ye Q, et al. Comparison of long read methods for sequencing and assembly of a plant genome. bioRxiv. 2020. https://doi.org/10.1101/2020.1103.1116.992933
17. Wimalanathan K, Lawrence-Dill CJ. Gene Ontology Meta Annotator for Plants. bioRxiv. 2019:809988.
18. Jung H, Jeon M-S, Hodgett M, Waterhouse P, Eyun S. A comparative evaluation of genome assemblers from long-read sequencing for plants and crops. J Agric Food Chem. 2020; 68:7670–7677. https://doi.org/10.1021/acs.jafc.0c01647 PMID: 32530283
19. Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, et al. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat Rev Genet. 2020:389–409.
20. Abdelrahman H, ElHady M, Alcivar-Warren A, Allen S, Al-Tobasei R, Bao L, et al. Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics. 2017; 18:191. https://doi.org/10.1186/s12864-017-3557-1 PMID: 28219347
21. Bernatchez L, Wellenreuther M, Araneda C, Ashton DT, Barth JMI, Beacham TD, et al. Harnessing the Power of Genomics to Secure the Future of Seafood. Trends Ecol Evol. 2017; 32:665–680. https://doi.org/10.1016/j.tree.2017.06.010 PMID: 28818341
22. Gratacap RL, Wargelius A, Edvardsen RB, Houston RD. Potential of Genome Editing to Improve Aquaculture Breeding and Production. Trends Genet. 2019; 35:672–684. https://doi.org/10.1016/j.tig.2019.06.006 PMID: 31331664
23. Shen Y, Yue G. Current status of research on aquaculture genetics and genomics-information from ISGA 2018. Aquaculture and Fisheries. 2019; 4:43–47.
24. Zenger KR, Khatkar MS, Jones DB, Khalilisamani N, Jerry DR, Raadsma HW. Genomic Selection in Aquaculture: Application, Limitations and Opportunities With Special Reference to Marine Shrimp and Pearl Oysters. Front Genet. 2018; 9:693. https://doi.org/10.3389/fgene.2018.00693 PMID: 30728827
25. Fan G, Song Y, Huang X, Yang L, Zhang S, Zhang M, et al. Initial data release and announcement of the Fish10K: Fish 10,000 Genomes Project. bioRxiv. 2019:787028.
26. Nguyen TV, Jung H, Rotllant G, Hurwood D, Mather P, Ventura T. Guidelines for RNA-seq projects: applications and opportunities in non-model decapod crustacean species. Hydrobiologia. 2018; 825:5–27.

27. Babarinde IA, Li Y, Hutchins AP. Computational Methods for Mapping, Assembly and Quantification for Coding and Non-coding Transcripts. Comput Struct Biotec. 2019; 17:628–637. https://doi.org/10.1016/j.csbj.2019.04.012 PMID: 31193391

28. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annu Rev Biomed Data Sci. 2019; 2:139–173.

29. Hölzer M, Marz M. *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience. 2019; 8:giz039. https://doi.org/10.1093/gigascience/giz039 PMID: 31077315

30. You X, Shan X, Shi Q. Research advances in the genomics and applications for molecular breeding of aquaculture animals. Aquaculture. 2020; 526:735357.

31. Pathak AK, Rashid I, Nagpure NS, Kumar R, Pati R, Singh M. FisOmics: A portal of fish genomic resources. Genomics. 2019; 111:1923–1928. https://doi.org/10.1016/j.ygeno.2019.01.003 PMID: 30611878

32. Rey C, Veber P, Boussau B, Semon M. CAARS: comparative assembly and annotation of RNA-Seq data. Bioinformatics. 2019; 35:2199–2207. https://doi.org/10.1093/bioinformatics/bty903 PMID: 30452539

33. Zhang X, Yuan J, Sun Y, Li S, Gao Y, Yu Y, et al. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat Commun. 2019; 10:356. https://doi.org/10.1038/s41467-018-08197-4 PMID: 30664654

34. Boivin V, Reulet G, Boisvert O, Couture S, Elela SA, Scott MS. Reducing the structure bias of RNA-Seq reveals a large number of non-annotated non-coding RNA. Nucleic Acids Res. 2020; 48:2271–2286. https://doi.org/10.1093/nar/gkaa028 PMID: 31980822

35. Nong W, Chai ZYH, Jiang X, Qin J, Ma KY, Chan KM, et al. A crustacean annotated transcriptome (CAT) database. BMC Genomics. 2020; 21:32. https://doi.org/10.1186/s12864-019-6433-3 PMID: 31918660

36. Tso CH, Wu JL, Lu MW. Blast2Fish: a reference-based annotation web tool for transcriptome analysis of non-model teleost fish. BMC Bioinformatics. 2020; 21:174. https://doi.org/10.1186/s12859-020-3507-9 PMID: 32366294

37. Zhu BH, Xiao J, Xue W, Xu GC, Sun MY, Li J-T. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 2018; 19:175. https://doi.org/10.1186/s12864-018-4567-3 PMID: 29499650

38. Gonzalez-Castellano I, Manfrin C, Pallavicini A, Martinez-Lage A. De novo gonad transcriptome analysis of the common littoral shrimp *Palaemon serratus*: novel insights into sex-related genes. BMC Genomics. 2019; 20:757.

39. Wang B, Kumar V, Olson A, Ware D. Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. Front Genet. 2019; 10:384.

40. Pootakham W, Uengwetwanit T, Sonthirod C, Sittikankaew K, Karoonuthaisiri N. A Novel Full-Length Transcriptome Resource for Black Tiger Shrimp (*Penaeus monodon*) Developed Using Isoform Sequencing (Iso-Seq). Front Mar Sci. 2020; 7:172.

41. Nguyen NH, Premachandra HKA, Kilian A, Knibb W. Genomic prediction using DArT-Seq technology for yellowtail kingfish *Seriola lalandi*. BMC Genomics. 2018; 19:107. https://doi.org/10.1186/s12864-018-4493-4 PMID: 29382299

42. Robledo D, Palaiokostas C, Bargelloni L, Martinez P, Houston R. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev Aquac. 2018; 10:670–682.

43. Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, et al. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2017; 17:142–152.

44. Matthews BJ, Vosshall LB. How to turn an organism into a model organism in 10 'easy' steps. J Exp Biol. 2020; 223:jeb218198. https://doi.org/10.1242/jeb.218198 PMID: 32034051

45. Kim BM, Amores A, Kang S, Ahn DH, Kim JH, Kim I-C, et al. Antarctic blackfin icefish genome reveals adaptations to extreme environments. Nat Ecol Evol. 2019; 3:469–478. https://doi.org/10.1038/s41559-019-0812-7 PMID: 30804520

46. Pettersson ME, Rochus CM, Han F, Chen J, Hill J, Wallerman O, et al. A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. Genome Res. 2019; 29:1919–1928. https://doi.org/10.1101/gr.253435.119 PMID: 31649060

47. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, da Veiga LF, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput Biol. 2016; 12:e1004947. https://doi.org/10.1371/journal.pcbi.1004947 PMID: 27415786

48. Carey MA, Papin JA. Ten simple rules for biologists learning to program. PLoS Comput Biol. 2018; 14: e1005871. https://doi.org/10.1371/journal.pcbi.1005871 PMID: 29300745

49. Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Pettersson OV, et al. Ten steps to get started in Genome Assembly and Annotation. F1000Research. 2018; 7:148. https://doi.org/10.12688/f1000research.13598.1 PMID: 29568489

50. Swathi A, Shekhar MS, Katneni VK, Vijayan KK. Genome size estimation of brackishwater fishes and penaeid shrimps by flow cytometry. Mol Biol Rep. 2018; 45:951–960. https://doi.org/10.1007/s11033-018-4243-3 PMID: 30008142

51. Fiske JA, Van Eenennaam JP, Todgham AE, Young SP, Holem-Bell CE, Goodbla AM, et al. A comparison of methods for determining ploidy in white sturgeon (*Acipenser transmontanus*). Aquaculture. 2019; 507:435–442.

52. Manekar SC, Sathe SR. Estimating the k-mer Coverage Frequencies in Genomic Datasets: A Comparative Assessment of the State-of-the-art. Curr Genomics. 2019; 20:2–15. https://doi.org/10.2174/1389202919666181026101326 PMID: 31015787

53. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020; 11:1432.

54. Pflug JM, Holmes VR, Burrus C, Spencer Johnston J, Maddison DR. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). bioRxiv. 2019:761304.

55. Hollenbeck CM, Johnston IA. Genomic Tools and Selective Breeding in Molluscs. Front Genet. 2018; 9:253. https://doi.org/10.3389/fgene.2018.00253 PMID: 30073016

56. Franěk R, Baloch AR, Kašpar V, Saito T, Fujimoto T, Arai K, et al. Isogenic lines in fish—a critical review. Rev Aquacult 2019. https://doi.org/10.1111/raq.12389

57. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS ONE. 2013; 8:e62856. https://doi.org/10.1371/journal.pone.0062856 PMID: 23638157

58. Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet. 2015; 16:627–640. https://doi.org/10.1038/nrg3933 PMID: 26442640

59. Sohn JI, Nam JW. The present and future of *de novo* whole-genome assembly. Brief Bioinform. 2018; 19:23–40. https://doi.org/10.1093/bib/bbw096 PMID: 27742661

60. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Brief Bioinform. 2019; 20:1542–1559. https://doi.org/10.1093/bib/bby017 PMID: 29617724

61. Wee Y, Bhyan SB, Liu Y, Lu J, Li X, Zhao M. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. Brief Funct Genomics. 2019; 18:1–12. https://doi.org/10.1093/bfgp/ely037 PMID: 30462154

62. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics. 2017; 18:474. https://doi.org/10.1186/s12859-017-1911-6 PMID: 29126390

63. Garg S, Rautiainen M, Novak AM, Garrison E, Durbin R, Marschall T. A graph-based approach to diploid genome assembly. Bioinformatics. 2018; 34:i105–i114. https://doi.org/10.1093/bioinformatics/bty279 PMID: 29949989

64. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. Genome Res. 2018; 28:1720–1732. https://doi.org/10.1101/gr.236273.118 PMID: 30341161

65. Jayakumar V, Sakakibara Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. Brief Bioinform. 2019; 20:866–876. https://doi.org/10.1093/bib/bbx147 PMID: 29112696

66. Tilak MK, Botero-Castro F, Galtier N, Nabholz B. Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. Genome Biol Evol. 2018; 10:616–622. https://doi.org/10.1093/gbe/evy022 PMID: 29385572

67. Wu WW, Phue JN, Lee CT, Lin C, Xu L, Wang R, et al. Robust Sub-nanomolar Library Preparation for High Throughput Next Generation Sequencing. BMC Genomics. 2018; 19:326. https://doi.org/10.1186/s12864-018-4677-y PMID: 29728062

68. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018; 34:666–681. https://doi.org/10.1016/j.tig.2018.05.008 PMID: 29941292

69. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2018; 19:1256–1272. https://doi.org/10.1093/bib/bbx062 PMID: 28637243

70. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. bioRxiv. 2019:735928.

71. Gaither MR, Gkafas GA, de Jong M, Sarigol F, Neat F, Regnier T, et al. Genomics of habitat choice and adaptive evolution in a deep-sea fish. Nat Ecol Evol. 2018; 2:680–687. https://doi.org/10.1038/s41559-018-0482-x PMID: 29507380

72. Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ, et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet. 2018; 50:270–277. https://doi.org/10.1038/s41588-017-0036-1 PMID: 29358652

73. Fu S, Wang A, Au KF. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol. 2019; 20:26.

74. Hench K, Vargas M, Hoppner MP, McMillan WO, Puebla O. Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. Nat Ecol Evol. 2019; 3:657–667. https://doi.org/10.1038/s41559-019-0814-5 PMID: 30833758

75. Wang K, Shen Y, Yang Y, Gan X, Liu G, Hu K, et al. Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. Nat Ecol Evol. 2019; 3:823–833. https://doi.org/10.1038/s41559-019-0864-8 PMID: 30988486

76. Ozerov MY, Ahmad F, Gross R, Pukk L, Kahar S, Kisand V, et al. Highly Continuous Genome Assembly of Eurasian Perch (*Perca fluviatilis*) Using Linked-Read Sequencing. G3. 2018; 8:3737–3743. https://doi.org/10.1534/g3.118.200768 PMID: 30355765

77. Dreau A, Venu V, Avdievich E, Gaspar L, Jones FC. Genome-wide recombination map construction from single individuals using linked-read sequencing. Nat Commun. 2019; 10:4309. https://doi.org/10.1038/s41467-019-12210-9 PMID: 31541091

78. Li C, Liu X, Liu B, Ma B, Liu F, Liu G, et al. Draft genome of the Peruvian scallop *Argopecten purpuratus*. GigaScience. 2018; 7:giy031. https://doi.org/10.1093/gigascience/giy031 PMID: 29617765

79. Louro B, De Moro G, Garcia C, Cox CJ, Verissimo A, Sabatino SJ, et al. A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*). GigaScience. 2019; 8:giy031.

80. Gong G, Dan C, Xiao S, Guo W, Huang P, Xiong Y, et al. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. GigaScience. 2018; 7:giy120.

81. Shao C, Li C, Wang N, Qin Y, Xu W, Liu Q, et al. Chromosome-level genome assembly of the spotted sea bass, *Lateolabrax maculatus*. GigaScience. 2018; 7:giy114. https://doi.org/10.1093/gigascience/giy114 PMID: 30239684

82. Bai CM, Xin LS, Rosani U, Wu B, Wang QC, Duan XK, et al. Chromosomal-level assembly of the blood clam, *Scapharca* (*Anadara*) *broughtonii*, using long sequence reads and Hi-C. GigaScience. 2019; 8:giz067. https://doi.org/10.1093/gigascience/giz067 PMID: 31289832

83. Xiao Y, Xiao Z, Ma D, Liu J, Li J. Genome sequence of the barred knifejaw *Oplegnathus fasciatus* (Temminck & Schlegel, 1844): the first chromosome-level draft genome in the family Oplegnathidae. GigaScience. 2019; 8:giz013. https://doi.org/10.1093/gigascience/giz013 PMID: 30715332

84. Endrullat C, Glokler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genom. 2016; 10:2–9. https://doi.org/10.1016/j.atg.2016.06.001 PMID: 27668169

85. Panova M, Aronsson H, Cameron RA, Dahl P, Godhe A, Lind U, et al. DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms. Methods Mol Biol. 2016; 1452:13–44. https://doi.org/10.1007/978-1-4939-3774-5_2 PMID: 27460368

86. Schiebelhut LM, Abboud SS, Gomez Daglio LE, Swift HF, Dawson MN. A comparison of DNA extraction methods for high-throughput DNA analyses. Mol Ecol Resour. 2017; 17:721–729. https://doi.org/10.1111/1755-0998.12620 PMID: 27768245

87. Arseneau JR, Steeves R, Laflamme M. Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. Mol Ecol Resour. 2017; 17:686–693. https://doi.org/10.1111/1755-0998.12616 PMID: 27768249

88. Cole BS, Moore JH. Eleven quick tips for architecting biomedical informatics workflows with cloud computing. PLoS Comput Biol. 2018; 14:e1005994. https://doi.org/10.1371/journal.pcbi.1005994 PMID: 29596416

89. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018; 19:208–219. https://doi.org/10.1038/nrg.2017.113 PMID: 29379135

90. Grossman RL. Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data. Trends Genet. 2019; 35:223–234.

91. Knapp B, Bardenet R, Bernabeu MO, Bordas R, Bruna M, Calderhead B, et al. Ten simple rules for a successful cross-disciplinary collaboration. PLoS Comput Biol. 2015; 11:e1004214. https://doi.org/10.1371/journal.pcbi.1004214 PMID: 25928184

92. Wei ZG, Zhang SW. NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model. BMC Bioinformatics. 2018; 19:177. https://doi.org/10.1186/s12859-018-2208-0 PMID: 29788930

93. Zhang W, Jia B, Wei C. PaSS: a sequencing simulator for PacBio sequencing. BMC Bioinformatics. 2019; 20:352.

94. Yue JX, Liti G. simuG: a general-purpose genome simulator. Bioinformatics. 2019; 35:4442–4444.

95. Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Comput Biol. 2015; 11:e1004191. https://doi.org/10.1371/journal.pcbi.1004191 PMID: 25905639

96. Chen P, Jing X, Ren J, Cao H, Hao P, Li X. Modelling BioNano optical data and simulation study of genome map assembly. Bioinformatics. 2018; 34:3966–3974. https://doi.org/10.1093/bioinformatics/bty456 PMID: 29893801

97. DeMaere MZ, Darling AE. Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies. GigaScience. 2018; 7:gix103. https://doi.org/10.1093/gigascience/gix103 PMID: 29149264

98. Li Y, Han R, Bi C, Li M, Wang S, Gao X. DeepSimulator: a deep simulator for Nanopore sequencing. Bioinformatics. 2018; 34:2899–2908. https://doi.org/10.1093/bioinformatics/bty223 PMID: 29659695

99. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017; 18:93. https://doi.org/10.1186/s13059-017-1213-3 PMID: 28521789

100. Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. BMC Genomics. 2017; 18:341. https://doi.org/10.1186/s12864-017-3723-5 PMID: 28464822

101. Eyun S, Soh HY, Posavi M, Munro J, Hughes DST, Murali SC, et al. Evolutionary history of chemosensory-related gene families across the Arthropoda. Mol Biol Evol. 2017; 34:1838–1862. https://doi.org/10.1093/molbev/msx147 PMID: 28460028

102. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013; 14:R47.

103. Yang LA, Chang YJ, Chen SH, Lin CY, Ho JM. SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies. BMC Genomics. 2019; 19:238.

104. Xu GC, Xu TJ, Zhu R, Zhang Y, Li SQ, Wang HW, et al. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. Gigascience. 2019;8. https://doi.org/10.1093/gigascience/giy157 PMID: 30576505

105. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 2018; 28:1029–1038. https://doi.org/10.1101/gr.233460.117 PMID: 29884752

106. Hosmani PS, Shippy T, Miller S, Benoit JB, Munoz-Torres M, Flores-Gonzalez M, et al. A quick guide for student-driven community genome annotation. PLoS Comput Biol. 2019; 15:e1006682. https://doi.org/10.1371/journal.pcbi.1006682 PMID: 30943207

107. Kong J, Huh S, Won JI, Yoon J, Kim B, Kim K. GAAP: A Genome Assembly + Annotation Pipeline. Biomed Res Int. 2019; 2019:4767354. https://doi.org/10.1155/2019/4767354 PMID: 31346518

108. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012; 13:329–342.

109. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. Nat Rev Genet. 2016; 17:758–772. https://doi.org/10.1038/nrg.2016.119 PMID: 27773922

110. Cruz F, Lagoa D, Mendes J, Rocha I, Ferreira EC, Rocha M, et al. SamPler—a novel method for selecting parameters for gene functional annotation routines. BMC Bioinformatics. 2019; 20:454. https://doi.org/10.1186/s12859-019-3038-4 PMID: 31488049

111. Wilbrandt J, Misof B, Panfilio KA, Niehuis O. Repertoire-wide gene structure analyses: a case study comparing automatically predicted and manually annotated gene models. BMC Genomics. 2019; 20:753. https://doi.org/10.1186/s12864-019-6064-8 PMID: 31623555

112. Cao H, Wahlestedt C, Kapranov P. Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. Trends Genet. 2018; 34:704–721.

113. Seetharam A, Singh U, Li J, Bhandary P, Arendsee Z, Wurtele ES. Maximizing prediction of orphan genes in assembled genomes. bioRxiv. 2019.

114. Permal E, Flutre T, Quesneville H. Roadmap for annotating transposable elements in eukaryote genomes. Methods Mol Biol. 2012; 859:53–68.

115. Wang Y, Chen L, Song N, Lei X. GASS: genome structural annotation for Eukaryotes based on species similarity. BMC Genomics. 2015; 16:150. https://doi.org/10.1186/s12864-015-1353-3 PMID: 25764973

116. König S, Romoth L, Stanke M. Comparative Genome Annotation. In: Setubal JC, Stoye J, Stadler PF, editors. Comparative Genomics: Methods and Protocols. New York, NY: Springer New York; 2018. pp. 189–212.

117. Jung J, Kim JI, Yi G. geneCo: a visualized comparative genomic method to analyze multiple genome structures. Bioinformatics. 2019; 35:5303–5305. https://doi.org/10.1093/bioinformatics/btz596 PMID: 31350879

118. Chowdhury B, Garai A, Garai G. An optimized approach for annotation of large eukaryotic genomic sequences using genetic algorithm. BMC Bioinformatics. 2017; 18:460. https://doi.org/10.1186/s12859-017-1874-7 PMID: 29065853

119. Jun S-R, Nookaew I, Hauser L, Gorin A. Assessment of genome annotation using gene function similarity within the gene neighborhood. BMC Bioinformatics. 2017; 18:345. https://doi.org/10.1186/s12859-017-1761-2 PMID: 28724412

120. Wilbrandt J, Misof B, Niehuis O. COGNATE: comparative gene annotation characterizer. BMC Genomics. 2017; 18:535. https://doi.org/10.1186/s12864-017-3870-8 PMID: 28716078

121. Geib SM, Hall B, Derego T, Bremer FT, Cannoles K, Sim SB. Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. Gigascience. 2018; 7:1–5.

122. Caballero M, Wegrzyn J. gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. Genomics Proteomics Bioinformatics. 2019; 17:305–310. https://doi.org/10.1016/j.gpb.2019.04.002 PMID: 31437583

123. Humann JL, Lee T, Ficklin S, Main D. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. Methods Mol Biol. 1962; 2019:29–51.

124. Brown AV, Campbell JD, Assefa T, Grant D, Nelson RT, Weeks NT, et al. Ten quick tips for sharing open genomic data. PLoS Comput Biol. 2018; 14:e1006472.

125. Boland MR, Karczewski KJ, Tatonetti NP. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. PLoS Comput Biol. 2017; 13:e1005278. https://doi.org/10.1371/journal.pcbi.1005278 PMID: 28103227

126. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol. 2015; 11:e1004525. https://doi.org/10.1371/journal.pcbi.1004525 PMID: 26492633

127. Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol. 2016; 12:e1005097.

128. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLoS Comput Biol. 2017; 13:e1005399. https://doi.org/10.1371/journal.pcbi.1005399 PMID: 28358831

129. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019; 15:e1006790. https://doi.org/10.1371/journal.pcbi.1006790 PMID: 30726205

130. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016; 17:66. https://doi.org/10.1186/s13059-016-0924-1 PMID: 27072794

131. Liu Y, Sargent L, Leung W, Elgin SCR, Goecks J. G-OnRamp: a Galaxy-based platform for collaborative annotation of eukaryotic genomes. Bioinformatics. 2019; 35:4422–4423. https://doi.org/10.1093/bioinformatics/btz309 PMID: 31070714

132. Sterck L, Billiau K, Abeel T, Rouze P, Van de Peer Y. ORCAE: online resource for community annotation of eukaryotes. Nat Methods. 2012; 9:1041. https://doi.org/10.1038/nmeth.2242 PMID: 23132114

133. Sargent L, Liu Y, Leung W, Mortimer NT, Lopatto D, Goecks J, et al. G-OnRamp: Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation. PLoS Comput Biol. 2020; 16:e1007863.

134. Long JC, Pomare C, Best S, Boughtwood T, North K, Ellis LA, et al. Building a learning community of Australian clinical genomics: a social network study of the Australian Genomic Health Alliance. BMC Med. 2019; 17:44.

135. Rexroad C, Vallet J, Matukumalli LK, Reecy J, Bickhart D, Blackburn H, et al. Genome to Phenome: Improving Animal Health, Production, and Well-Being—A New USDA Blueprint for Animal Genome Research 2018–2027. Front Genet. 2019; 10:327. https://doi.org/10.3389/fgene.2019.00327 PMID: 31156693

136. Stark Z, Boughtwood T, Phillips P, Christodoulou J, Hansen DP, Braithwaite J, et al. Australian Genomics: A Federated Model for Integrating Genomics into Healthcare. Am J Hum Genet. 2019; 105:7–14. https://doi.org/10.1016/j.ajhg.2019.06.003 PMID: 31271757

137. Pedro H, Yates AD, Kersey PJ, De Silva NH. Collaborative Annotation Redefines Gene Sets for Crucial Phytopathogens. Front Microbiol. 2019; 10:2477.

Evolutionary Applications
Open Access
Evolutionary Applications ISSN 1752-4571

REVIEWS AND SYNTHESIS

A field guide to whole-genome sequencing, assembly and annotation

Robert Ekblom and Jochen B. W. Wolf

Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

Keywords
bioinformatics, conservation genomics, genome assembly, next generation sequencing, vertebrates, whole - genome sequencing.

Correspondence
Robert Ekblom, Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, SE- 752 36 Uppsala, Sweden.
Tel.: +46 18 471 6468;
fax: +46 18 471 6310;
e-mail: robert.ekblom@ebc.uu.se

Received: 4 February 2014
Accepted: 20 May 2014

doi:10.1111/eva.12178

Abstract

Genome sequencing projects were long confined to biomedical model organisms and required the concerted effort of large consortia. Rapid progress in high-throughput sequencing technology and the simultaneous development of bioinformatic tools have democratized the field. It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate *de novo* draft genome sequences for any organism of choice. Because of the cost and considerable effort involved in such an endeavour, the important first step is to thoroughly consider whether a genome sequence is necessary for addressing the biological question at hand. Once this decision is taken, a genome project requires careful planning with respect to the organism involved and the intended quality of the genome draft. Here, we briefly review the state of the art within this field and provide a step-by-step introduction to the workflow involved in genome sequencing, assembly and annotation with particular reference to large and complex genomes. This tutorial is targeted at scientists with a background in conservation genetics, but more generally, provides useful practical guidance for researchers engaging in whole-genome sequencing projects.

Introduction

The field of conservation genetics is concerned with studying genetic and evolutionary processes in the context of biodiversity conservation (Frankham et al. 2010). Traditionally, a small number of neutral genetic markers were employed to study patterns of genetic variation of individuals and populations with the aim to explore underlying processes and their relevance to conservation. Marker-based measures provide insight into effective population sizes (N_e), recent demographic events (e.g. bottlenecks and expansions), genetic relatedness and the level of inbreeding. Genetic markers are also routinely employed in monitoring schemes (e.g. identification of individuals and capture–recapture modelling) and in breeding programs (Romanov et al. 2009) and have been extensively used to characterize population substructuring and genetic connectivity, to delineate conservation units and to infer interspecific admixture events (Höglund 2009; Allendorf et al. 2013). Under the premise of a causal relationship between neutral genetic variation and population viability, these data can inform practical conservation decisions (Frankham 1995).

Rapid advances in sequencing technology and bioinformatic tools during the last decade have initiated a transition from classical conservation genetics to conservation genomics (Fig. 1; Primmer 2009; Allendorf et al. 2010; Ouborg et al. 2010; Steiner et al. 2013). This development has two major implications. First, by significantly scaling-up the number of genetic markers, genomewide approaches enhance the power and resolution for the above-mentioned applications and improve the reliability of conclusions (Steiner et al. 2013). Second, the application of genomic technologies opens novel axes of investigation (Allendorf et al. 2010; Ouborg et al. 2010). Genome-scale data provide information beyond neutral genetic variation or candidate gene approaches (e.g. major histocompatibility complex genes; Hedrick 1999) and thus enable screening for selectively important variation and assessing the adaptive potential of populations (Primmer 2009). For example, approaches such as genomewide scans for selection, association mapping or quantitative trait loci (QTL) mapping can pinpoint loci of relevance for local adaptation of the target population (Steiner et al. 2013), with the potential to conserve

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

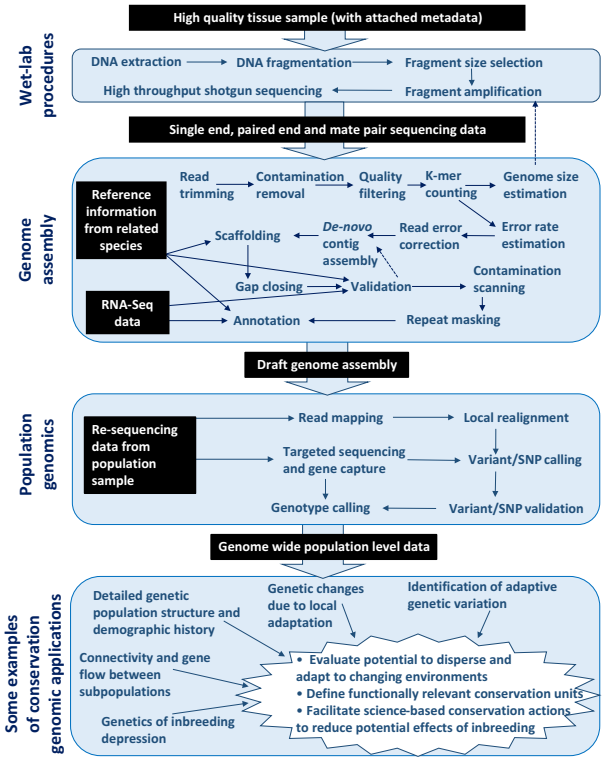


Figure 1 Workflow of a typical *de novo* whole-genome sequencing project. Black boxes with white text indicate genomic resources becoming available during the course of the project. From the top: wet-lab procedures, *de novo* assembly bioinformatic pipeline, postassembly analyses of additional population-wide sampling (population genomics), conservation genomic questions to address and analyses to perform (conservation genomic applications). Bullet points within the white star in the bottom part of the figures represent ultimate goals in conservation biology that can be addressed using genomic information combined with high-quality ecological data.

evolutionary processes – a long sought after goal in conservation biology (Crandall et al. 2000; Fraser and Bernatchez 2001). Genomewide analyses further allow addressing the poorly understood mechanistic basis of inbreeding depression (epistasis, directional dominance versus overdominance, many versus few loci), or assessing the impact of genetic variation on patterns of gene expression, and plastic response to environmental change. Genomic approaches can also be applied to highly fragmented DNA from ancient material (e.g. from museum specimens; Pääbo et al. 2004; Bi et al. 2013), to characterize environmental samples (Shokralla et al. 2012) and to understand how environmental perturbations affect microbial communities (Mardis 2008), representing largely unexplored terrain in conservation biology.

The above-mentioned applications do not necessarily require a reference genome sequence. Many analyses, including taxonomic delineation, characterization of demographic events, estimates of inbreeding or relatedness, can be successfully conducted in the absence of a genome reference. Instead, large-scale maker data such as genotyping-by-sequencing (Elshire et al. 2011), RAD-Seq (Narum et al. 2013), reduced representation sequencing (Van Tassell et al. 2008), amplicon sequencing (Zavodna et al. 2013) or transcriptome sequencing (Ekblom and Galindo 2011) can be effectively utilized without relying on a genome backbone. A complete and well-annotated genome sequence, however, provides the ultimate resource for genomic approaches. Whole-genome resequencing data with positional information along a genome sequence constitute the most complete account of individual genomic variation [e.g. structural rearrangements, copy number variation, insertion–deletion, single nucleotide polymorphisms (SNPs), sequence repeats] and will likely soon become the standard for genetic studies of natural populations (Ellegren et al. 2012; The Heliconius Genome Consortium 2012). It also provides the basis for haplotype information and genomewide estimates of linkage disequilibrium which have great power to reveal recent population histories (Li and Durbin 2011), timing of admixture events (Hellenthal et al. 2014) and to screen for signatures of selection (Hohenlohe et al. 2010). The study of selectively important variation strongly relies on annotated genome data to identify the functional genomic regions of interest. Reference sequences are further indispensable as a template for RNA-seq in detailed studies of (isoform-specific, allele-specific) gene expression (Vijay et al. 2013), epigenetic modifications (such as methylation; Herrera and Bazaga 2011) and DNA–protein interactions (Auerbach et al. 2013). These approaches are only accessible to genome-enabled taxa (Kohn et al. 2006) that enjoy the added benefit of using the latest bioinformatics tools developed in the biomedical sciences.

Here, we introduce the workflow of a typical whole-genome sequencing project conducted by an individual research group. This field guide aims at introducing principles and concepts to beginners in the field (Box 1) and offers practical guidance for the many steps involved (Fig. 1). It builds largely upon our own experience with vertebrate genome assembly. We limit the scope to genomic data, focusing on large and complex genomes, for transcriptome assembly we refer to Martin and Wang (2011) and Wolf (2013). We discuss sequencing, assembly and annotation, highlighting typical routines and analytical procedures. Our intention is not to provide a comprehensive review of sequencing technology, assembly algorithms or downstream downstream analyses, as this has already been performed. For these topics, we instead list exemplary literature and provide relevant entry points (Box 2).

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

Box 1: Glossary

Alignment	Similarity-based arrangement of DNA, RNA or protein sequences. In this context, subject and query sequence should be orthologous and reflect evolutionary, not functional or structural relationships
Annotation	Computational process of attaching biologically relevant information to genome sequence data
Assembly	Computational reconstruction of a longer sequence from smaller sequence reads
Barcode	Short-sequence identifier for individual labelling (barcoding) of sequencing libraries
BAC	(Bacterial artificial chromosome) DNA construct of various length (150–350 kb)
cDNA	Complementary DNA synthesized from an mRNA template
Contig	A contiguous linear stretch of DNA or RNA consensus sequence. Constructed from a number of smaller, partially overlapping, sequence fragments (reads)
Coverage	Also known as 'sequencing depth'. *Sequence coverage* refers to the average number of reads per locus and differs from *physical coverage*, a term often used in genome assembly referring to the cumulative length of reads or read pairs expressed as a multiple of genome size
De novo assembly	Refers to the reconstruction of contiguous sequences without making use of any reference sequence
EST library	Expressed sequence tag library. A short subsequence of cDNA transcript sequence
Fosmid	A vector for bacterial cloning of genomic DNA fragments that usually holds inserts of around 40 kb
GC content	The proportion of guanine and cytosine bases in a DNA/RNA sequence
Gene ontology (GO)	Structured, controlled vocabularies and classifications of gene function across species and research areas
InDel	Insertion/deletion polymorphism
Insert size	Length of randomly sheared fragments (from the genome or transcriptome) sequenced from both ends
K-mer	Short, unique element of DNA sequence of length k, used by many assembly algorithms
Library	Collection of DNA (or RNA) fragments modified in a way that is appropriate for downstream analyses, such as high-throughput sequencing in this case
Mapping	A term routinely used to describe alignment of short sequence reads to a longer reference sequence
Masking	Converting a DNA sequence [A,C,G,T] (usually repetitive or of low quality) to the uninformative character state N or to lower case characters [a,c,g,t] (*soft masking*)
Massively parallel (or next generation) sequencing	High-throughput sequencing nano-technology used to determine the base-pair sequence of DNA/RNA molecules at much larger quantities than previous end-termination (e.g. Sanger sequencing) based sequencing techniques
Mate-pair	Sequence information from two ends of a DNA fragment, usually several thousand base-pairs long
N50	A statistic of a set of contigs (or scaffolds). It is defined as the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs
N90	Equivalent to the N50 statistic describing the length for which the collection of all contigs of that length or longer contains at least 90% of the total of the lengths of the contigs
Optical map	Genomewide, ordered, high-resolution restriction map derived from single, stained DNA molecules. It can be used to improve a genome assembly by matching it to the genomewide pattern of expected restriction sites, as inferred from the genome sequence
Paired-end sequencing	Sequence information from two ends of a short DNA fragment, usually a few hundred base pairs long
Read	Short base-pair sequence inferred from the DNA/RNA template by sequencing
RNA-Seq	High-throughput shotgun transcriptome (cDNA) sequencing. Usually not used synonymous to RNA-sequencing which implies direct sequencing of RNA molecules skipping the cDNA generation step
Scaffold	Two or more contigs joined together using read-pair information
Transcriptome	Set of all RNA molecules transcribed from a DNA template

Basic considerations

Genome assembly is a challenging problem that requires time, resources and expertise. Before engaging in a genome sequencing project, it should thus be carefully considered whether a genome reference sequence is strictly necessary for the purpose in question. Genome sequences are merely a resource and in many cases will contribute very little *per se* to a problem in conservation biology. In case a genome draft is judged to be of significant value to address the problem at hand, it needs to be considered whether sufficient financial and computational resources are available to produce a genome of satisfactory quality. If funding is not available to obtain the appropriate read depth, it is advisable to utilize alternative approaches where possible (such as genotyping-by-sequencing or transcriptome sequenc-

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

Box 2: Before you start

Some important points to consider

- Availability of appropriate computational resources
- Collaboration with sequencing facility and bioinformatics groups
- Plan for amount and type of sequencing data needed
- Does funding allow to produce sufficient sequence coverage? If not, alternative approaches should be considered rather than producing a poor, low coverage, assembly
- Familiarization with data handling pipelines and file formats (see below)
- High-quality DNA sample (with individual metadata)
- Plan for analyses and publication

Some useful resources

Internet forums for discussions related to genome sequencing

- http://seqanswers.com/
- http://www.biostars.org/
- http://www.biosupport.se/

Entry points to genome sequencing, assembly and exemplary downstream analyses

- Library preparation and Sequencing: Mardis (2008, 2013)
- Quality filtering/preprocessing: Patel and Jain (2012), Zhou and Rokas (2014), Smeds and Künstner (2011)
- Genome assembly: Nagarajan and Pop (2013), Pop (2009), Flicek and Birney (2009)
- Assembly evaluation: Earl et al. (2011), Bradnam et al. (2013), Bao et al. (2011)
- Genome annotation: Yandell and Ence (2012)
- Mapping: Li and Durbin (2009), Trapnell and Salzberg (2009), Bao et al. (2011)
- Data handling: Li et al. (2009), Quinlan and Hall (2010)
- Variant calling: Nielsen et al. (2011), DePristo et al. (2011), Van der Auwera et al. (2013)
- Haplotype-based approaches: Browning and Browning (2011), Tewhey et al. (2011), Lawson et al. (2012)
- Population genomic summary statistics: Nielsen et al. (2012b), Danecek et al. (2011)

Web resources

- Galaxy (http://galaxyproject.org/)
- Amazon cloud (http://aws.amazon.com/ec2/)
- Windows Azure (http://www.windowsazure.com/)
- Magellan: Cloud Computing for Science (http://www.alcf.anl.gov/magellan)
- Web Apollo (http://genomearchitect.org/)
- NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject/)
- Genomes OnLine Database (http://genomesonline.org/cgi-bin/GOLD/index.cgi)
- ENSEMBL genome database (http://www.ensembl.org/index.html)
- UCSC Genome Browser (http://genomebrowser.wustl.edu/)
- fastQCtoolkit for data preprocessing (http://www.bioinformatics.babraham.ac.uk/projects/fastqc)

Genome size databases

- Plants: http://data.kew.org/cvalues/
- Animals: http://www.genomesize.com/

Common file formats

• FASTA	Nucleotide sequence (file extension .fas or .fa)
• FASTQ	Nucleotide sequence including quality scores
• SAM	Sequence alignment
• BAM	Binary version of SAM
• GFF3	Annotation
• GTF	Annotation
• BED	Annotation
• VCF	Variant calling

ing), rather than settle for low-coverage whole-genome sequencing data. The latter would be a waste of funding, effort and time.

One important limitation of the current shotgun genome sequencing approaches that may be of particular importance in conservation biology is the fact that core genes with high conservation relevance, like immune genes of the MHC or olfactory receptor (OR) genes, are highly polymorphic and have many paralogs, which makes them particularly difficult to assemble. More generally, rapidly evolving genes or members of large gene families are often poorly represented in the final assembly and annotated gene set. Such regions and genes constitute a challenge even for very large sequencing projects of model organisms. If the project is not carefully planned from the start, there is a risk that the regions of highest interest to conservation biology will not be correctly represented in the final draft of the genome. Manual annotation and use of additional data, such as targeted sequencing of bacterial artificial chromosome (BAC) clones, will often be necessary to include such genomic regions in the assembly. If information on such preidentified candidate genes is the main aim of the study, it might even be more efficient to focus only on those regions rather than trying to sequence and assemble the whole-genome (see for example Wang et al. 2012).

What does it mean to 'sequence a genome'?

Ideally, a genome draft would represent the complete nucleotide base sequence for all chromosomes in the species of interest, a 'physical map' of its genetic content (as opposed to the 'genetic or linkage map' which establishes the order and recombination distances among genetic markers). However, in reality, there are a number of complications with the concept of a 'genome sequence'. First, there is not one true sequence for a species because of individual genomic variation. In a single diploid individual, such variation will manifest itself in the form of heterozygous positions, insertion/deletion (InDel) polymorphism, copy number variation or small-scale rearrangements. Even cells from the same individual can differ in genomic content due to somatic mutations. The assembled genome sequence of an individual will also be only one representation of the total variation present in a species (paralleling

the use of 'type specimens' for taxonomic classification). Generally, only a single individual is sequenced (Wheeler et al. 2008), but sometimes (like in the HUGO project) the genome represents a 'consensus' of a number of pooled samples (International Human Genome Sequencing Consortium 2004). Note, however, that in diploid and polyploid organisms, the genome assembly already reflects a consensus sequence of several chromosome sets and fails to capture haplotypic variation (for most current short-read based methods). Second, it is essentially impossible to sequence and assemble all nucleotides in the genome (Ellegren 2014). Large parts of DNA sequence, especially the heterochromatic regions around centromeres and telomeres and other highly repetitive regions, are not well-characterized even in mature genome assemblies like human or mice. Third, there will always be some degree of error in the characterized genome sequence, both on the level of individual nucleotides (stemming from sequencing errors) and in the ordering of sequence blocks (stemming from assembly errors). Forth, every genome assembly is the result of a series of assembly heuristics and should accordingly be treated as a working hypothesis.

The principle of genome sequencing and assembly

Currently, most genome projects use a shotgun sequencing strategy for genome sequencing (Fig. 2). In a first step, genomic DNA is sheared into small random fragments. Depending on the technology, these are sequenced independently to a given length. Powerful computer algorithms are then utilized to piece the resulting sequence reads back together into longer continuous stretches of sequence (*contigs*), a process known as *de novo* assembly. For correct assembly, it is important that there is sufficient overlap between the sequence reads at each position in the genome, which requires high sequencing coverage (or read depth). Naturally, for longer sequence reads, more overlap can be expected, reducing the required raw read depth. Usually, longer fragments (several hundred base pairs) are sequenced from both ends (paired-end sequencing) to provide additional information on correct read placement in the assembly.

After the initial assembly, *contigs* are typically joined to form longer stretches of sequence (known as *scaffolds*). To achieve this, libraries from long DNA fragments spanning several kilobases (kb) of sequence in the genome are prepared and their endpoints sequenced. Depending on the technology and the specifics of the library preparation, these libraries are (somewhat confusingly) called, for example paired-end, mate-pair or jump libraries. If the endpoint sequences of several independent fragments come to lie on two different *contigs*, they are joined into a *scaffold*. The expected fragment length of the library provides information on the physical distance between the two *contigs*, and the created gap is filled with the uninformative base-pair character 'N'. Subsequent gap closing methods, ideally using long reads that read across repetitive sequences, help to fill in the missing base-pair information.

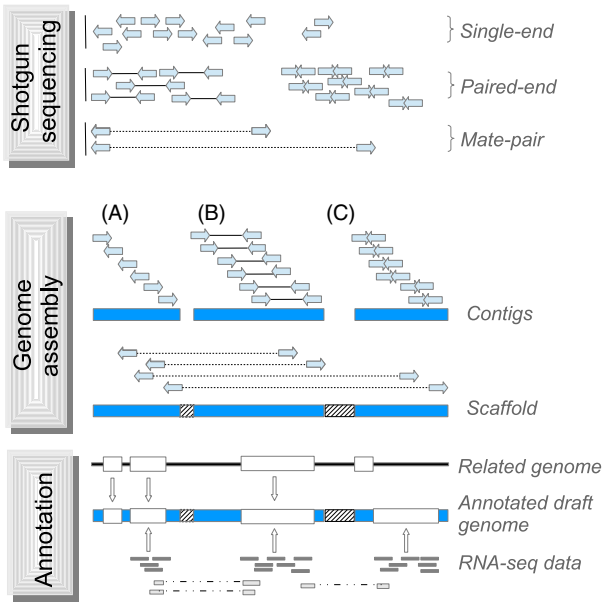


Figure 2 Simplified illustration of the assembly process and terminology. Shotgun sequencing: short fragments of DNA from the target organism are sequenced at random positions across the genome to a given depth of coverage. Fragments can consist of single reads (typically 50–1000 bp) or of paired-end reads of varying insert size (note that paired-end reads can even overlap). Mate-pair libraries span larger genomic regions (~2–20 kb inserts) with reads generally facing outwards and can be complemented with fosmid-end libraries (~40 kb inserts). Genome assembly: (A) short-read *de novo* assemblers extend the disperse sequence information from the reads into continuous stretches called *contigs*. *Contigs* usually reflect the consensus sequence and do not contain any polymorphisms. (B) Paired-end reads provide additional information on whether a read is supported for a given *contig*. (C) Some assemblers such as ALLPATHS-LG work with overlapping read pairs that are joined into a virtual longer read prior to the assembly. Read pairs from mate-pair or fosmid-end libraries can be used to order and orient *contigs* into *scaffolds*. Gap size between *contigs* is estimated from the expected length of mate-pairs and marked with 'N's (indicated by hatched grey boxes). Long reads from single molecule sequencing provide an alternative. Annotation: gene models can be inferred *in silico* by prediction algorithms, by lifting over information from genomes of related organisms and by using transcriptome data (RNA-seq, expressed sequence tag) from the target organism itself. Spliced reads from RNA-seq data as indicated at the bottom of the figure provide valuable evidence for splice junctions and various isoforms of a gene.

In a last step, the resulting *scaffolds* are often joined into linkage groups or placed on chromosomes (Ellegren 2014). Genetic maps constructed from pedigree data or crosses are arguably the best way for ordering and orienting *scaffolds* into longer sequence blocks (Ellegren et al. 2012). However,

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

detailed genetic maps of species with conservation concern (usually not amendable to artificial crosses or half-sib breeding designs) require substantial genotyping effort, and deep pedigrees with a sufficient number of meioses are difficult to come by in most systems (Romanov et al. 2009). Given these difficulties, it is often not realistic to aim for a chromosome-level assembly, and this will also often not be necessary for most conservation biology applications. Most applications, including haplotype-based approaches that are powerful in revealing signatures of selection or depict recent demographic histories, generally work with high-quality *contigs*. As an alternative for placing and orienting the *scaffolds* onto putative chromosomes, synteny and gene order information from related species can be used. Note, however, that such information should be used with due caution as chromosomal rearrangements may have occurred even between very closely related species. There is also a risk that errors in the reference species assembly are transferred to the focal genome.

Genome sequencing

Sequencing technology and coverage considerations

Among the first decisions when starting, a genome sequencing project is the choice of sequencing platform, the type and amount of sequence data to generate. The latter is often limited by project funding, and the former may depend on which sequencing technology is promptly available. Judging from recently completed whole-genome sequencing projects (Table 1), there is a clear trend moving away from traditional Sanger sequencing (~1 kb sequence reads) and Roche 454 sequencing (up to 800 bp) towards short read technologies such as Illumina HiSeq (at present typically 150 bp) and SOLiD (typically 50 bp). Lately, there has been progress in producing longer reads at high throughput; several technologies offering this, such as Pacific Biosciences (up to 5 kb), IonTorrent (~500 bp) and Illumina Moleculo (up to 10 kb), are entering the market, and we expect to see a broader spectrum of read lengths. While this development blurs the initial dichotomy of short reads (e.g. 35 bp Illumina reads) versus long reads (~1 kb Sanger reads), read length still has important bioinformatic implications, as assembly algorithms optimized for long reads are fundamentally different from approaches targeting short reads. Recent studies begin to combine data of different read length and from several different sequencing platforms (Koren et al. 2012). This strategy makes intuitive sense as the drawbacks of each method can be counterbalanced, although the jury is still out whether such hybrid assemblies always outperform single data type approaches (Bradnam et al. 2013). Here, we follow the principle of current common practice and base our considerations largely on sequencing of Illumina libraries of different lengths (we loosely refer to short reads at sequence lengths below 500 bp and long reads above this length). Many of the following reflections, however, more generally relate to the assembly problem and do not depend on the specific choice of sequencing library.

For most downstream applications, obtaining long *contigs* is essential. With long-read data, from traditional Sanger sequencing of individual BAC clones, this is feasible even with a rather limited sequencing depth. However, when using only short-read technologies, high total read coverage (>100×) is needed. Too little data will result in a highly fragmented assembly and severe problems with downstream applications such as annotation and variant calling. For initial *contig* assembly, one usually starts out with a high amount of paired-end short-read data. To subsequently merge *contigs* into *scaffolds*, it is necessary to generate additional libraries with long-insert sizes in the range of 3–40 kb (Fig. 2). How much sequencing data are needed of each library types and insert size depends critically on a number of factors including the size and repeat content of the genome, the degree of heterozygosity and the target quality of the assembly (Sims et al. 2014). As these parameters will differ between sequencing projects and organisms of interest, the optimal resource allocation will be unique to every project. As a rough guideline for mammalian genomes, it has been proposed to use at least 45× coverage of short-insert paired-end libraries, 45× coverage of medium-sized insert libraries (3–10 kb) and 1–5× coverage of long-insert libraries (10–40 kb) (Nagarajan and Pop 2013). It should be noted that coverage can sometimes also be too high, as the absolute number of sequencing errors increases as a function of read number. According to our own experience, down-sampling from 100× to 50× coverage of a short-insert size library can significantly improve some steps in the assembly process.

To translate these recommendations into amount and type of sequencing needed for a specific project, basic knowledge on genome size, sequencing error rates, repeat content and the degree of genome duplications is needed. If no such information is available for the target species of interest at the start of the project, it is advisable to first perform a small pilot study using single-end or short-insert sequencing. The above-mentioned parameters can then be approximated using a k-mer counting approach (Marçais and Kingsford 2011; http://josephryan.github.com/estimate_genome_size.pl). Information on how to perform and interpret such k-mer counts can be found in web forums such as seqanswers (Box 2). Generally, a larger amount of long-insert data is needed for correct assembly if the genome has a high repeat content or a high degree of duplications. Genome size estimates for a large number of species are also available in online databases (Box 2).

Table 1. Some recently sequenced vertebrate genomes in species of conservation concern.

Species	Red list category	Sequencing technology	Assembly algorithm	Contig N50 (bp)	Sequencing coverage	Number of authors	References
Chimpanzee	EN	Sanger	PCAP	53000	6×	67	Consortium (2005)
Mammoth	EX	Roche 454	NA	NA	<1×	22	Miller et al. (2008)
Panda	EN	Illumina GA	SOAPdenovo	39886	56×	123	Li et al. (2010)
Orang-utan	CR	Sanger	PCAP	15654	6×	101	Locke et al. (2011)
Cod	VU	Roche 454	Newbler	2778	40×	42	Star et al. (2011)
Tasmanian devil	EN	Roche 454/Illumina GAIIx	Newbler/CABOG	9495	14×	30	Miller et al. (2011)
African elephant	VU	Sanger (ABI3730)	ARACHNE (reference assisted)	2900	2×	60	Lindblad-Toh et al. (2011)
Tarsier	NT	Sanger (ABI3730)	ARACHNE (reference assisted)	2900	2×	60	Lindblad-Toh et al. (2011)
Polar bear	VU	Illumina HiSeq 2000	SOAPdenovo	3596	100×	26	Miller et al. (2012)
Puerto Rican parrot	CR	Illumina HiSeq 2000	Ray	6983	27×	14	Oleksyk et al. (2012)
Gorilla	CR	Sanger/Illumina	Phusion assembler/ABySS	11800	50×	71	Scally et al. (2012)
Bonobo	EN	Roche 454	Celera Assembler	67000	25×	41	Prufer et al. (2012)
Yak	VU	Illumina HiSeq 2000	SOAPdenovo	20400	65×	48	Qiu et al. (2012)
Aye-aye	NT	Illumina GAIIx	CLC bio Assemler	3650	38×	10	Perry et al. (2012)
Coelacanth	CR	Illumina HiSeq 2000	ALLPATHS-LG	12700	61×	91	Amemiya et al. (2013)
Saker falcon	EN	Illumina HiSeq 2000	SOAPdenovo	31200	113×	25	Zhan et al. (2013)
Tibetan antelope	EN	Illumina GAIIx	SOAPdenovo (reference assisted)	NA	Not reported	11	Kim et al. (2013)
Bluefin tuna	LC*	Roche 454/Illumina GAIIx	Newbler/Bowtie	7588	54×	24	Nakamura et al. (2013)
Darwin's finch	LC†	Roche 454	Newbler	Not reported	4×	19	Rands et al. (2013)
Straw coloured fruit bat	NT	Illumina HiSeq 2000	CLC bio	27140	17×	7	Parker et al. (2013)
King cobra	VU	Illumina GAIIx	CLC/SSPACE	3980	40×	36	Vonk et al. (2013)
Burmese python	VU	Roche 454/Illumina HiSeq 2000	Newbler/SOAPdenovo	10700	49×	39	Castoe et al. (2013)
Chinese softshell turtle	VU	Illumina HiSeq 2000	SOAPdenovo	22000	106×	34	Wang et al. (2013)
Tiger	EN	Illumina HiSeq 2000	SOAPdenovo	29800	118×	58	Cho et al. (2013)
Minke whale	LC‡	Illumina HiSeq 2000	SOAPdenovo	22571	128×	55	Yim et al. (2014)
Northern bobwhite	NT	Illumina HiSeq 2000	CLC	45400	142×	12	Halley et al. (2014)
Black grouse	LC§	SOLiD 5500xl	SOAPdenovo (reference assisted)	1238	127×	5	Wang et al. (2014)
White rhinoceros	NT	Illumina HiSeq 2000	ALLPATHS-LG	93000	91×	10	Di Palma et al. unpublished data

Red list categories: EX, extinct; CR, critically endangered; EN, endangered; VU, vulnerable, NT, near threatened; LC, least concern.
*Not red-listed, but likely to be affected by overfishing.
†Not red-listed, but endemic to a small geographic region.
‡Not currently red-listed but, subject to extensive exploitation or within group of endangered taxa.
§Not globally red-listed, but with several small and isolated regional populations.

Wet-lab procedures

The wet-lab part of the genome sequencing is often outsourced to sequencing centres, and we will only very briefly touch upon the basic steps of library preparation that are important to consider at the planning stage of the project and that affect downstream analytical procedures.

Genome individual

Heterozygous positions in the genome of the sequenced individual have adverse effects on the assembly. Highly polyploid species are particularly challenging for assembly and may necessitate specifically tailored assembly pipelines (Schatz et al. 2012). A general recommendation is to use inbred individuals, parthenogenetic or gynogenetic off-

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

spring where available. Attached to the genome individual should be metadata that might be important for future referencing, such as the identity, age and sex of the individual, time and exact place of sampling, etc. (Genome 10K Community of Scientists 2009).

Tissue

Energetically active tissue (such as muscle) should be avoided, as there is a risk that the sequence data will contain a high proportion of mitochondrial DNA (mtDNA), which wastes sequencing effort and can cause problems in the assembly step due to the extreme read depth (as assembly pipelines often use read depth to identify duplicated genomic regions). We further recommend removing mtDNA sequence reads prior to assembly and use only a small fraction of this data to assemble the mitochondrial genome (which in itself may provide important information for conservation genetics) separately. It is also advisable to avoid tissues such as gut and skin which may have a high degree of nontarget DNA from xenobiotic organisms.

DNA quality

Whole-genome sequencing, particularly of long-insert size libraries, requires high-quality, intact, nondegraded DNA at a sufficient amount (Wong et al. 2012). For sequencing, a full genome using a set of different libraries requires ~1 mg of DNA as starting material (~6 μg for short-insert libraries, ~40 μg for 2–10 kb libraries, ~60 μg for >20 kb libraries). Before engaging in genome sequencing, it is thus essential to obtain a large amount of high-quality DNA of the target species. This can be a major obstacle for many species with conservation concern. If captive animals are available, such samples can often be utilized as a source of high-quality DNA, but note that genomic variation identified from such sources may not be representative of wild populations. Prior to submitting a DNA sample, its integrity should be checked on a high-resolution gel (e.g. pulse-field electrophoresis; a sample should typically show fragments of >100 kb).

Library preparation

When choosing the necessary raw read depth, one should be aware that currently most technologies include several PCR steps which can lead to a non negligible number of duplicated reads. While single reads can occur in duplicate by chance if coverage is high enough, duplication is bound to be an artefact for identical read pairs which are very unlikely to occur by chance (as they follow a length distribution). As duplicated reads are of no added value and duplication artefacts can impair coverage-based quality validation, they should be removed prior to the assembly. Duplicates generally constitute a few percentage of short-insert size libraries (<500 bp), but can reach over 95% for long-insert libraries (>10 kb).

Another central question refers to what insert sizes to use. Generally, it is advisable to have a good mix of sizes in the range of 0.2–40 kb with the shorter libraries being sequenced to significantly higher depth (Gnerre et al. 2011). Insert sizes of >20 kb make a large difference to the final contiguity and *scaffold* size of the assembly, but are not trivial to produce at high quality and currently constitute a limitation of many sequencing centres. Library preparations differ in quality and in how well they represent different parts of the genome. Therefore, more than one library should ideally be generated per size class. Note that some assembly programs (such as ALLPATHS-LG) expect a predefined mix of sequencing libraries as input data. Another important issue for downstream analyses that comes with library preparation is read orientation. Depending on the technology used, reads can face inwards (→ ←; e.g. Illumina paired-end sequencing) or outward (← →; e.g. Illumina mate-pair sequencing) in relation to the original DNA fragment. Mis-oriented reads with unexpectedly short insert sizes can arise due to sequencing of pairs from within the original DNA fragment rather than at its ends. Also, mate-pairs with aberrant insert sizes and orientation often represent chimeric sequences from nonadjacent genomic regions. For most assembly methods, such artefacts need to be filtered out during the preassembly steps, often leaving only a small fraction of usable, unique read pairs for assembly after trimming. To correctly process the data, the bioinformatician handling the data always needs to be 'library aware'.

Genome assembly

Data management

The amounts of data generated in a normal genome sequencing project is staggering. A vertebrate genome with 100× coverage means data files in the order of several hundred gigabytes. During the assembly procedure, temporary files easily cross the terabyte boundary. An adequate data management and backup strategy is thus needed already at the start of the project. Many universities are connected to local or national computing grids, including data storage facilities, and it is highly recommended to utilize these whenever possible. Having bioinformatics application experts working at the computing infrastructure provides a vital link between the biologist researchers and the computing grid system experts (Lampa et al. 2013). Such collaborations should already be established during the planning stage of the project. Sequencing centres often also offer assistance with data analyses and assembly. However, their automated pipelines are not likely to be optimized for data from nonmodel organisms and might not be usable from a conservation biology point of view. It is thus vital to explicitly discuss what kind of support can be

© 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

provided by the facility before the start of the project. More generally, it should be considered whether enough expertise exists in the core research group to perform the computational steps of an assembly. Most data processing and genomic analyses of large-scale sequencing data are conducted on high-performance computing clusters running a UNIX-based operating system. One does not need to be a bioinformatics expert to handle whole-genome sequencing data, but is essential to have some familiarity with the UNIX environment and basic knowledge in command line software, writing shell scripts and applying scripts of commonly used languages for biological data analysis (such as Perl or Python).

Preassembly steps

Prior to the assembly, the quality of the sequencing data, overall GC content, repeat abundance or the proportion of duplicated reads should be assessed. Tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) providing summary statistics are a useful starting point. Trimming low-quality data and reads resulting from PCR duplications can be performed with a variety of different software and scripts (e.g. ConDeTri; Smeds and Künstner 2011). Stand-alone error correcting, using a k-mer count approach (for example as applied in the SOAPdenovo pipeline), can also be a useful alternative for many datasets. Note, however, that the optimal stringency of quality filtering is specific to the individual project and the targeted assembly pipeline. Some assemblers, such as ALLPATHS-LG (Gnerre et al. 2011), where trimming and error correction are performed within the assembly pipeline, even require raw reads, without quality trimming as input.

Primer and vector sequences from the library preparation will most likely be present in the data (even if the sequencing facility claims to have removed them) and can be removed with simple scripts (like cutadapt; Martin 2011). Also, in Illumina sequencing, DNA from the PhiX phage is often added to the sequencing reaction, in order to calibrate sequence quality scores. Failure to remove such abundant contaminant sequences can disrupt the assembly process (due to the high read depth compared with the nuclear genome) and may result in the production of chimeric and contaminated *contigs*. The easiest way of removing known vector contamination from the raw data is to use a short read aligner (like BWA; Li and Durbin 2009) and delete all fragments mapping to the contamination sequence.

De novo assembly

Tools for genome assembly differ widely in their performance in terms of speed, scalability and the quality of the final genome sequence (Miller et al. 2010; Earl et al. 2011; Narzisi and Mishra 2011; Bradnam et al. 2013). While some assembly methods clearly outperform others, it is currently difficult to predict which of the tools might be most appropriate in a given situation. Every assembly project is unique in terms of generated data structure and the target genome differing, for instance, in size, base-composition, repeat content and polymorphism level. There are a number of software available for *de novo* assembly of shotgun whole-genome sequencing data, and new programs are constantly being added to the list. Some algorithms focus on minimizing mis-assemblies, while others mainly aim to improve contiguity (sometimes at the cost of accuracy). Most assembly algorithms perform optimally with a given distribution of library sizes, so it is important to consider the choice of assembly strategy already during the project planning and sequencing steps. Besides information from the primary literature and websites of assembly software, various web forums provide good entry points for up-to-date discussions and sharing the experiences of other researchers (see Box 2).

Most software implementations designed for long-read technologies such as traditional Sanger sequencing (for example the Celera assembler, Arachne and PCAP; Batzoglou et al. 2002; Huang et al. 2003; Denisov et al. 2008) or Roche 454 sequencing (for example Newbler) use an assembly approach known as overlap-layout-consensus (OLC). These algorithms are generally considered too computationally intensive (mainly in terms of runtime) for Illumina or SOLiD data. Still, a few assemblers such as Edena (Hernandez et al. 2008), SGA (Simpson and Durbin 2012) and FERMI (Li 2012) pursue the OLC strategy for such short-read data (Miller et al. 2010). Most other strategies for *de novo* assembly of short sequence reads can be broadly divided into two classes: extension-based methods and De Bruijn (or Eulerian) graph algorithms (Nagarajan and Pop 2013). Extension-based assemblers, such as SSAKE (Warren et al. 2007) and JR-Assembler (Chu et al. 2013) are usually computationally very efficient (in terms of both memory requirements and computational time), but are highly sensitive to sequencing errors, repeat regions and high levels of nucleotide polymorphism (Chu et al. 2013). The most commonly used approach for assembly of short-read data is therefore currently based on De Bruijn graphs, where reads are partitioned into k-mers (substrings of the read sequence of length k) that then form the nodes of the graph (network) and are linked when sharing a k-1 mer. Highly used assembly software, such as SOAPdenovo (Luo et al. 2012), ALLPATHS-LG (Gnerre et al. 2011), ABySS (Simpson et al. 2009) and Velvet (Zerbino and Birney 2008), all rely on De Bruijn graph algorithms. There are also 'hybrid' assembly approaches, for example Atlas (Havlak et al. 2004), Ray (Boisvert et al. 2010) and MaSuRCA

 © 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

(Zimin et al. 2013), combining features of different algorithms and utilizing data from multiple sequencing technologies. In general, it is advisable to test several different assembly methods and evaluate which is most appropriate for the specific data at hand. Draft genome building should be treated as an iterative process with several rounds of assembly, evaluation and parameter tweaking. For a more comprehensive review of different assembly algorithms and software; see for example (Miller et al. 2010; Nagarajan and Pop 2013).

After the initial *contig* building, it is common to use read-pair information from long-insert (mate-pair, fosmid-end or jump) libraries (Zhang et al. 2012) to combine *contigs* into *scaffolds*. Additional short-insert paired-end libraries are also often useful, for example to bridge, short low-complexity regions. The lengths of sequence gaps between *contigs* are estimated from the expected insert sizes and are usually filled with a stretch of Ns. The scaffolding step is already included in many commonly used assembly programs, but there are also some stand-alone applications, for example SSPACE (Boetzer et al. 2011) and BESST (Nystedt et al. 2013), to perform this step independently. Some of the gaps (N's) emerging from this process can be removed a posteriori using the original read-pair information with software such as GapCloser (Li et al. 2010), GapFiller (Boetzer and Pirovano 2012) and iMAGE (Tsai et al. 2010). Long-read data (for example from PacBio) has also recently emerged as a way of filling N regions in *scaffolds* (English et al. 2012).

When choosing assembly software, it is important to consider both the amount of sequencing data and which computational resources are available (Schatz et al. 2010a). De Bruijn graph methods, such as SOAPdenovo and ALLPATHS-LG, generally require large amounts of computing memory (RAM). Depending on the amount of sequencing data, assembly of mammal-sized genomes (~3 Gb) can require terabytes of internal memory (Lampa et al. 2013). If large computer clusters are not available locally, it will be necessary to consider collaborative equipment purchases, joint projects with bioinformatics groups or utilization of commercially available computing clouds (Box 2; Schatz et al. 2010b).

Another consideration to make is whether to use freely available programs (most programs mentioned above fall in this category) or to invest in commercial software (such as CLC workbench or Lasergene from DNASTAR). Commercial software is usually more user-friendly than freely available programs and thus readily used by researchers with limited bioinformatics skills. The downside of commercial software, apart from the (often substantial) cost involved in purchase and licensing, is that these act even more like 'black box' solutions, where it is often near impossible to inspect or alter details about the algorithms. Some commonly used software applications are also distributed together with the sequencing instruments and may be available through the sequencing facilities.

For an increasing number of species with conservation concern, there are genomic information available for very closely related taxa (Kohn et al. 2006). In such cases, it is an attractive alternative to use as much information as possible from the related genome in the assembly process (Gnerre et al. 2009). Such a 'reference-assisted' assembly strategy has been utilized in several studies (Table 1), mainly using custom pipelines, and there is clearly great scope for development of more mature software in this field. The most common approach is to first produce *contigs de novo* and then align these to the genomic reference to aid in the scaffolding step. Assuming extensive synteny and gene order conservation, such an approach makes it possible to build large *scaffolds* even with low coverage data (Kim et al. 2013) or using very short sequence reads (Wang et al. 2014). An alternative approach is the so-called 'Align-Layout-Consensus' algorithm. Here, the overlap stage of the *de novo* assembly is replaced by alignment of reads to a closely related reference genome (which is computationally less intensive compared with the OLC approach). *Contigs* and *scaffolds* are then built *de novo*, using information from overlapping reads (Schneeberger et al. 2011).

Quality assessment and validation

Once an assembly has been successfully performed, users will want to assess its quality or compare several assemblies using different methods. Yet, as discussed above, every draft genome assembly constitutes merely a hypothesis of the true underlying genome sequence, and in the absence of knowing the truth, assessing its quality remains a challenge.

A variety of metrics reflecting different aspects of the assembly are available (Bradnam et al. 2013). They can be broadly divided into approaches that require additional information from external data and those solely based on information derived from the assembly itself. As external information is often not available in conservation genomics projects, intrinsic quality assessment of the assembly is a natural starting point. One basic metric is the proportion of the genome contained within the assembly. The expected genome size can either be inferred from *C*-value data (Box 2) or, alternatively, from k-mer frequency-based approaches. Another standard metric to evaluate assembly contiguity is the N50 statistic: by definition, 50% of the assembled nucleotides are found in *contigs* (contig N50) or *scaffolds* (scaffold N50) of at least this length. The N50 statistic thus describes a kind of median of assembled sequence lengths, giving greater weight to long sequences.

Recently, variations of this metric, for example the NG50 and 'NG Graph' (Bradnam et al. 2013), incorporating the expected genome size was introduced and provide effective means of visualizing and evaluating differences in contiguity between assemblies.

However, the N50 statistic and variations thereof need to be interpreted with caution. They merely indicate contiguity and contain no information on assembly accuracy. To detect errors in the assembly, information from remapped paired-end or mate-pair data can be used (as, e.g., implemented in the software REAPR; Hunt et al. 2013). Low-coverage regions or mis-orientation of read pairs suggests mis-assemblies, while aberrant insert sizes indicate small insertions or deletions. Exceedingly high sequence coverage, high local SNP densities or correlated SNPs, where most of the reads carry one character state (but multiple others show another character state), can indicate the presence of collapsed, near-identical repeats. Software applications performing these steps are numerous, and examples can be found in the current literature (Earl et al. 2011; Bradnam et al. 2013). The *amosvalidate* pipeline (Phillippy et al. 2008; Schatz et al. 2013) encompasses several genome assembly diagnostics in one pipeline, but works best for small- or medium-sized genomes.

Independent experimental data sets from the target species arguably provide the best source of external information. Data from optical maps, for instance, allow validating short- and large- scale accuracy of *scaffolds* and expanding them further to approach chromosome level. Similarly, separately assembled sequences from BAC or fosmid libraries can help to assess sequence accuracy and repeat content. Both approaches, however, rely on correct assembly themselves and are not readily available for smaller laboratories at present. Independent *de novo* assemblies from shotgun transcriptome sequencing data (RNA-seq) are more easily generated, and expressed sequence tag (EST) libraries might already exist for species of conservation concern (although getting access to fresh tissues for RNA extraction may be a serious limitation if captive populations are not available). Sequence content and exon structure of transcriptome data thus constitute an important additional resource for validating sequence accuracy and for correcting scaffolding in cases where genes span across *contigs*.

Comparative genomic approaches provide another avenue, which does not require the generation of additional data. For example, quantifying the presence and completeness of orthologous core eukaryotic protein sequences (Parra et al. 2007) provides first intuition on the comprehensiveness of the assembly. In cases where high-quality reference genomes of sister taxa exist, genome comparisons might be of guidance in detecting mis-assemblies and chimeric *contigs* under the assumption of broad-scale synteny and gene order conservation. Small-scale rearrangements, however, might be real and require in depth investigation.

DNA from other organisms are likely to have contaminated the genomic samples at various stages (during both sampling and laboratory procedures) and will be present in the sequencing data. Although mainly being a nuisance, contaminations at the sampling stage may actually be interesting from a conservation point of view, as they can carry information about parasites or other microorganisms related to the study species. External genomic resources aid in finding such contaminations that might have been assembled as separate *contigs* or are interspersed with target sequence in the same *contig*. Positive hits from a BLAST search or similar local alignment routines are often employed to find such traces of contamination, but results need to be interpreted with caution. Even correctly assembled sequences can lead to best hits from distantly related species with well-annotated genomes, particularly if taxon sampling within the target group of organisms is scarce. Likewise, small stretches of contamination in a large *contig* or *scaffold* may be missed entirely if other parts of the sequence yield significant hits on the target clade. Human contamination in other mammalian genome sequences will be particularly problematic, as such contamination is expected to be common due to handling of the samples. For parts of a *de novo*-sequenced mammalian genome, the best BLAST hit will be against a human or mouse sequence simply because the region in question has not been sequenced and annotated in any other mammal.

Genome annotation

To harness the full potential of a genome sequence, it needs to be annotated with biologically relevant information that can range from gene models and functional information, such as gene ontology (GO) terms (Gene Ontology Consortium 2004; Primmer et al. 2013) or 'Kyoto encyclopedia of genes and genomes' (KEGG) pathways (Kanehisa and Goto 2000), to microRNA and epigenetic modifications (The ENCODE Project Consortium 2012). In the context of genetic nonmodel organisms, annotation is often confined to protein-coding sequence (CDS) or transcripts more generally. Despite the considerable challenge to annotate genes in newly sequenced species where preexisting gene models are mostly lacking, automated gene annotation has in principle become possible for individual research groups (Yandell and Ence 2012). Still, a complete genome annotation constitutes a considerable effort and requires bioinformatic proficiency. We describe only the general workflow and refer the interested reader to a comprehensive review by Yandell and Ence (2012) for more details (Box 2). Before starting, it should be noted that

 © 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

successful annotation strongly depends on the quality of the genome assembly. Only contiguous near-complete (~90%) genomes interrupted only by small gaps will yield satisfying results. As a rule of thumb, large genomes have longer genes and thus need more contiguous assemblies for successful annotation (cf. Figure 1 in Yandell and Ence 2012).

The annotation process can be conceptually divided into two phases: a 'computational phase' where several lines of evidence from other genomes or from species-specific transcriptome data are used in parallel to create initial gene and transcript predictions. In a second 'annotation phase', all (sometimes contradicting) information is then synthesized into a gene annotation, following a set of rules determined by the annotation pipeline.

Prior to gene prediction, it is of vital importance to mask repetitive sequences including low-complexity regions and transposable elements. As repeats are often poorly conserved across species, it is advisable to create a species-specific repeat library using tools like RepeatModeler or RepeatExplorer (Novák et al. 2013). Once repeats are masked (e.g. with RepeatMasker; http://www.repeatmasker.org), *ab initio* algorithms trained on gene models from related species can be used for baseline prediction of coding sequence (CDS) (e.g. AUGUSTUS; Stanke et al. 2006). Protein alignments (using e.g. tblastx) and syntenic protein lift-overs from a variety of other species provide a valuable resource to complement the predicted gene models. Arguably, the best evidence comes from detailed EST or RNA-seq data, which in addition to CDS, provides gene models with information on splice sites, transcription start sites and untranslated regions (UTRs). If possible, mRNA should be sequenced strand-specifically, as this helps resolve gene models, facilitates transcriptome assembly and eventually aids in the evaluation of the genome assembly.

In a next step, all the evidence from *ab initio* prediction and protein-, EST- or RNA-alignments need to be synthesized into a final set of gene annotations. As the evidence is mostly incomplete and sometimes contradicting, this is a difficult task that often benefits from manual curation. Still, several automated annotation tools like MAKER (Cantarel et al. 2008) or PASA (Haas et al. 2003) exist that incorporate, and weigh the evidence from, several sources. Although these tools generally provide good results, qualitative validation is important (e.g. by assessing the length of open-reading frames). Visual inspection of the annotation is another vital component to detect systematic issues such as intron leakage (introns being annotated as exons due to the presence of pre-mRNA) or gene fusion. Tools like WebApollo (Lee et al. 2013) from the GMOD project are particularly useful, as they allow the user to edit the annotation directly through the visual interface.

Publishing the genome

Draft genome sequences are now being produced at an ever-increasing rate. Traditional databases such as ENSEMBL from the European Molecular Biology Labs (EMBL) and the Wellcome Trust Sanger Institute, or genomic databases from the National Center for Biotechnology Information (NCBI) providing access to genomes and meta-information can no longer annotate and curate all incoming genomes. NCBI therefore already provides the possibility to upload draft genome sequences and user-generated annotation. To allow other users to improve the assembly and its annotation, all available raw data should be uploaded, together with the assembled genome and all relevant meta-data, for example as a BioProject on NCBI.

Perspectives

Conservation applications

We have summarized information on current methods for whole-genome sequencing, assembly and annotation, with the aim of providing practical guidance for conservation or ecology-oriented research groups moving into the field of genomics. The focus has been on large and complex genomes of nonmodel organisms relevant from a conservation perspective. In the introduction, we outlined a number different ways in which genomic resources in general, and a complete genome sequence, in particular, can be applied in a conservation biology setting (see also Fig. 1). Conservation genomics being a young field, examples where genomic resources have been put to the test in an applied conservation context are still limited, but a few such cases may be worth highlighting.

One of the first nonmodel genomes to be sequenced using the Illumina technology was the giant panda (Li et al. 2010). While the focus of the panda genome paper was not on conservation issues, follow-up studies have utilized the draft genome to make inferences about population structure, adaptive genetic variation and demography (Wei et al. 2012). Likewise in the Aye-Aye, resequencing data from twelve individuals from different parts of Madagascar were utilized to infer fine-scale genetic population structure and conduct landscape genetic analyses. The results were used to provide guidance for allocation of conservation resources towards preserving large and contiguous habitats in northern Madagascar (Perry et al. 2013). Genomic resources have further been utilized in breeding programs of the Tasmanian devil, which is endangered in the wild due to a contagious facial cancer. The generation of a reference genome sequence in combination with genomewide resequencing data has made it possible to investigate many details of this disease, including the identification of candidate genes involved in tumorigenesis (Murchison et al.

2012). Similarly, genomic resources have been utilized to limit the spread of a developmental disease causing mutation in breeding programs of the California condor (Romanov et al. 2009). Finally, genomewide SNP screening has been effective in several studies of fishery stock monitoring and management (Primmer 2009; Nielsen et al. 2012a).

Future directions

With rapid progress in sequencing nano-technology and further development of computational methods, we can expect that all steps of the workflow will continue to be improved. New library preparation protocols will enable sequencing from less starting material, producing libraries with longer and more precisely estimated insert sizes and generating longer reads with reduced error rates. The development of more efficient assembly algorithms and increasing computational power will make the bioinformatic data processing amenable to a larger spectrum of users. As the costs involved in genome sequencing and assembly continues to drop, the generation of a draft genome sequence will soon become routine, also for species with large genomes. This development will mean that even small research groups with limited funding will soon be expected to develop genomic resources for their species of choice, reinforcing the use of genomic approaches in conservation biology and related disciplines. The possible development of rapid and compact sequencing solutions that may be applied directly in the field situation would be particularly useful for many conservation applications. Another important area of progress lies in the usage of low-quality samples, obtained from noninvasive sampling or museum material that would allow monitoring of genomic diversity through time. Developing ways of storing and sharing genomic data will also be crucial, to make the most efficient use of these resources for conservation. In spite of these promising developments, we need to be aware that science alone is not sufficient to meet future conservation challenges. The technical transition from conservation genetics to genome-scale data therefore needs to be tightly accompanied by a discussion of how applied conservation biology can best benefit from genomic data (see for example McMahon et al., 2014). This discussion needs to be taken at the general level on a case-by-case basis and involves scientists and political decision makers alike.

Acknowledgements

We thank the organisers and participants of the 'Evolutionary conservation' symposium at ESEB 2013 for discussions, and Christophe Eizaguirre and Miguel Soares for inviting this contribution. Douglas Scofield, Aaron Shafer and three anonymous reviewers provided feedback on earlier drafts of this manuscript. We also wish to thank all participants of the EBC Next Generation Sequencing Journal Club (Uppsala University) for stimulating discussions on this subject over several years. RE was funded by the Swedish Environmental Protection Agency [grant number 235-12-11].

Literature cited

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics **11**:697–709.

Allendorf, F. W., G. H. Luikart, and S. N. Aitken 2013. Conservation and the Genetics of Populations, 2nd edn. Wiley-Blackwell, Chichester.

Amemiya, C. T., J. Alfoldi, A. P. Lee, S. Fan, H. Philippe, I. MacCallum, I. Braasch et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature **496**:311–316.

Auerbach, R. K., B. Chen, and A. J. Butte 2013. Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. Bioinformatics **29**:1922–1924.

Bao, S., R. Jiang, W. Kwan, B. Wang, X. Ma, and Y.-Q. Song 2011. Evaluation of next-generation sequencing software in mapping and assembly. Journal of Human Genetics **56**:406–414.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger et al. 2002. ARACHNE: a whole-genome shotgun assembler. Genome Research **12**:177–189.

Bi, K., T. Linderoth, D. Vanderpool, J. M. Good, R. Nielsen, and C. Moritz 2013. Unlocking the vault: next-generation museum population genomics. Molecular Ecology **22**:6018–6032.

Boetzer, M., and W. Pirovano 2012. Toward almost closed genomes with GapFiller. Genome Biology **13**:R56.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**:578–579.

Boisvert, S., F. Laviolette, and J. Corbeil 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of Computational Biology **17**:1519–1533.

Bradnam, K., J. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience **2**:10.

Browning, S. R., and B. L. Browning 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**:703–714.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research **18**:188–196.

Castoe, T. A., A. P. J. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R. P. Ruggiero et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proceedings of the National Academy of Sciences **110**:20645–20650.

The Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69–87.

Cho, Y. S., L. Hu, H. Hou, H. Lee, J. Xu, S. Kwon, S. Oh et al. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. Nature Communications **4**:Article number: 2433.

Chu, T.-C., C.-H. Lu, T. Liu, G. C. Lee, W.-H. Li, and A. C.-C. Shih 2013. Assembler for de novo assembly of large genomes. Proceedings of the National Academy of Sciences **110**:E3417–E3424.

 © 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

Crandall, K. A., O. R. P. Bininda-Emonds, G. M. Mace, and R. K. Wayne 2000. Considering evolutionary processes in conservation biology. Trends in Ecology & Evolution **15**:290–295.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker et al. 2011. The variant call format and VCFtools. Bioinformatics **27**:2156–2158.

Denisov, G., B. Walenz, A. L. Halpern, J. Miller, N. Axelrod, S. Levy, and G. Sutton 2008. Consensus generation and variant detection by Celera Assembler. Bioinformatics **24**:1035–1040.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics **43**:491–498.

Earl, D., K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu et al. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Research **21**:2224–2241.

Ekblom, R., and J. Galindo 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity **107**:1–15.

Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution **29**:51–63.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backstrom, T. Kawakami, A. Kunstner et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature **491**:756–760.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**: e19379.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7**:e47768.

Flicek, P., and E. Birney 2009. Sense from sequence reads: methods for alignment and assembly. Nature Methods **6**(11 Suppl):S6–S12.

Frankham, R. 1995. Conservation genetics. Annual Review of Genetics **29**:305–327.

Frankham, R., J. D. Ballou, and D. A. Briscoe 2010. Introduction to Conservation Genetics, 2nd edn. Cambridge University Press, Cambridge, UK.

Fraser, D. J., and L. Bernatchez 2001. Adaptive evolutionary conservation: towards a unified concept for defining conservation units. Molecular Ecology **10**:2741–2752.

Gene Ontology Consortium 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research **32**(Suppl 1):D258–D261.

Genome 10K Community of Scientists 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. Journal of Heredity **100**:659–674.

Gnerre, S., E. Lander, K. Lindblad-Toh, and D. Jaffe 2009. Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biology **10**:R88.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences **108**:1513–1518.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr, L. I. Hannick, R. Maiti et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research **31**:5654–5666.

Halley, Y. A., S. E. Dowd, J. E. Decker, P. M. Seabury, E. Bhattarai, C. D. Johnson, D. Rollins et al. 2014. A draft *de novo* genome assembly for the northern bobwhite (*Colinus virginianus*) reveals evidence for a rapid decline in effective population size beginning in the Late Pleistocene. PLoS One **9**:e90240.

Havlak, P., R. Chen, K. J. Durbin, A. Egan, Y. Ren, X.-Z. Song, G. M. Weinstock et al. 2004. The Atlas genome assembly system. Genome Research **14**:721–732.

Hedrick, P. W. 1999. Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution **53**:313–318.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers 2014. A genetic atlas of human admixture history. Science **343**:747–751.

Hernandez, D., P. François, L. Farinelli, M. Østerås, and J. Schrenzel 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Research **18**:802–809.

Herrera, C. M., and P. Bazaga 2011. Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. Molecular Ecology **20**:1675–1688.

Höglund, J. 2009. Evolutionary Conservation Genetics. Oxford University Press, Oxford.

Hohenlohe, P. A., P. C. Phillips, and W. A. Cresko 2010. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. International Journal of Plant Sciences **171**:1059–1071.

Huang, X., J. Wang, S. Aluru, S.-P. Yang, and L. Hillier 2003. PCAP: a whole-genome assembly program. Genome Research **13**:2164–2170.

Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. Otto 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biology **14**:R47.

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. Nature **431**:931–945.

Kanehisa, M., and S. Goto 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**:27–30.

Kim, J., D. M. Larkin, Q. Cai, Asan, Y. Zhang, R.-L. Ge, L. Auvil et al. 2013. Reference-assisted chromosome assembly. Proceedings of the National Academy of Sciences **110**:1785–1790.

Kohn, M. H., W. J. Murphy, E. A. Ostrander, and R. K. Wayne 2006. Genomics and conservation genetics. Trends in Ecology & Evolution **21**:629–637.

Koren, S., M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology **30**:693–700.

Lampa, S., M. Dahlo, P. Olason, J. Hagberg, and O. Spjuth 2013. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. GigaScience **2**:9.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush 2012. Inference of population structure using dense haplotype data. PLoS Genetics **8**: e1002453.

Lee, E., G. Helt, J. Reese, M. C. Munoz-Torres, C. Childers, R. M. Buels, L. Stein et al. 2013. Web Apollo: a web-based genomic annotation editing platform. Genome Biology **14**:R93.

Li, H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics **28**:1838–1844.

Li, H., and R. Durbin 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760.

Li, H., and R. Durbin 2011. Inference of human population history from individual whole-genome sequences. Nature **475**:493–496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078–2079.

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang et al. 2010. The sequence and de novo assembly of the giant panda genome. Nature **463**:311–317.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature **478**:476–482.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature **469**:529–533.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**:18.

Marçais, G., and C. Kingsford 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764–770.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics **9**:387–402.

Mardis, E. R. 2013. Next-generation sequencing platforms. Annual Review of Analytical Chemistry **6**:287–303.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal **17**:10–12.

Martin, J. A., and Z. Wang 2011. Next-generation transcriptome assembly. Nature Reviews Genetics **12**:671–682.

McMahon, B. J., E. C. Teeling, and J. Höglund. 2014. How and why should we implement genomics into conservation? Evolutionary Applications **7**:999–1007.

Miller, W., D. I. Drautz, A. Ratan, B. Pusey, J. Qi, A. M. Lesk, L. P. Tomsho et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature **456**:387–390.

Miller, J. R., S. Koren, and G. Sutton 2010. Assembly algorithms for next-generation sequencing data. Genomics **95**:315–327.

Miller, W., V. M. Hayes, A. Ratan, D. C. Petersen, N. E. Wittekindt, J. Miller, B. Walenz et al. 2011. Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil). Proceedings of the National Academy of Sciences **108**:12348–12353.

Miller, W., S. C. Schuster, A. J. Welch, A. Ratan, O. C. Bedoya-Reina, F. Zhao, H. L. Kim et al. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proceedings of the National Academy of Sciences **109**:E2382–E2390.

Murchison, E. P., O. B. Schulz-Trieglaff, Z. Ning, L. B. Alexandrov, M. J. Bauer, B. Fu, M. Hims et al. 2012. Genome sequencing and analysis of the tasmanian devil and its transmissible cancer. Cell **148**:780–791.

Nagarajan, N., and M. Pop 2013. Sequence assembly demystified. Nature Reviews Genetics **14**:157–167.

Nakamura, Y., K. Mori, K. Saitoh, K. Oshima, M. Mekuchi, T. Sugaya, Y. Shigenobu et al. 2013. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. Proceedings of the National Academy of Sciences **110**:11061–11066.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. A. Hohenlohe 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology **22**:2841–2847.

Narzisi, G., and B. Mishra 2011. Comparing de novo genome assembly: the long and short of it. PLoS One **6**:e19175.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics **12**:443–451.

Nielsen, E. E., A. Cariani, E. M. Aoidh, G. E. Maes, I. Milano, R. Ogden, M. Taylor et al. 2012a. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature Communications **3**:851.

Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang 2012b. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One **7**:e37558.

Novák, P., P. Neumann, J. Pech, J. Steinhaisl, and J. Macas 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics **29**:792–793.

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin, D. G. Scofield, F. Vezzi et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature **497**:579–584.

Oleksyk, T., J.-F. Pombert, D. Siu, A. Mazo-Vargas, B. Ramos, W. Guiblet, Y. Afanador et al. 2012. A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education. GigaScience **1**:14.

Ouborg, N. J., C. Pertoldi, V. Loeschcke, R. Bijlsma, and P. W. Hedrick 2010. Conservation genetics in transition to conservation genomics. Trends in Genetics **26**:177–187.

Pääbo, S., H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch et al. 2004. Genetic analyses from ancient DNA. Annual review of genetics **38**:645–679.

Parker, J., G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka, and S. J. Rossiter 2013. Genome-wide signatures of convergent evolution in echolocating mammals. Nature **502**:228–231.

Parra, G., K. Bradnam, and I. Korf 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061–1067.

Patel, R. K., and M. Jain 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One **7**:e30619.

Perry, G. H., D. Reeves, P. Melsted, A. Ratan, W. Miller, K. Michelini, E. E. Louis et al. 2012. A genome sequence resource for the aye-aye (*Daubentonia madagascariensis*), a nocturnal lemur from Madagascar. Genome Biology and Evolution **4**:126–135.

Perry, G. H., E. E. Louis, A. Ratan, O. C. Bedoya-Reina, R. C. Burhans, R. Lei, S. E. Johnson et al. 2013. Aye-aye population genomic analyses highlight an important center of endemism in northern Madagascar. Proceedings of the National Academy of Sciences **110**:5823–5828.

Phillippy, A., M. Schatz, and M. Pop 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biology **9**:R55.

Pop, M. 2009. Genome assembly reborn: recent computational challenges. Briefings in Bioinformatics **10**:354–366.

Primmer, C. R. 2009. From conservation genetics to conservation genomics. Annals of the New York Academy of Sciences **1162**:357–368.

Primmer, C. R., S. Papakostas, E. H. Leder, M. J. Davis, and M. A. Ragan 2013. Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. Molecular Ecology **22**:3216–3241.

Prufer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature **486**:527–531.

Qiu, Q., G. Zhang, T. Ma, W. Qian, J. Wang, Z. Ye, C. Cao et al. 2012. The yak genome and adaptation to life at high altitude. Nature Genetics **44**:946–949.

Quinlan, A. R., and I. M. Hall 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841–842.

 © 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

Rands, C., A. Darling, M. Fujita, L. Kong, M. Webster, C. Clabaut, R. Emes et al. 2013. Insights into the evolution of Darwin's finches from comparative analysis of the Geospiza magnirostris genome sequence. BMC Genomics **14**:95.

Romanov, M. N., E. M. Tuttle, M. L. Houck, W. S. Modi, L. G. Chemnick, M. L. Korody, E. M. S. Mork et al. 2009. The value of avian genomics to the conservation of wildlife. BMC Genomics **10**(Suppl 2): S10.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature **483**:169–175.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg 2010a. Assembly of large genomes using second-generation sequencing. Genome Research **20**:1165–1173.

Schatz, M. C., B. Langmead, and S. L. Salzberg 2010b. Cloud computing and the DNA data race. Nature Biotechnology **28**:691.

Schatz, M., J. Witkowski, and W. R. McCombie 2012. Current challenges in de novo plant genome sequencing and assembly. Genome Biology **13**:243.

Schatz, M. C., A. M. Phillippy, D. D. Sommer, A. L. Delcher, D. Puiu, G. Narzisi, S. L. Salzberg et al. 2013. Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Briefings in Bioinformatics **14**:213–224.

Schneeberger, K., S. Ossowski, F. Ott, J. D. Klein, X. Wang, C. Lanz, L. M. Smith et al. 2011. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proceedings of the National Academy of Sciences **108**:10249–10254.

Shokralla, S., J. L. Spall, J. F. Gibson, and M. Hajibabaei 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology **21**:1794–1805.

Simpson, J. T., and R. Durbin 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Research **22**:549–556.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and İ. Birol 2009. ABySS: a parallel assembler for short read sequence data. Genome Research **19**:1117–1123.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics **15**:121–132.

Smeds, L., and A. Künstner 2011. ConDeTri – a content dependent read trimmer for Illumina data. PLoS One **6**:e26314.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research **34**(Suppl 2):W435–W439.

Star, B., A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrom, T. F. Gregers, T. B. Rounge et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature **477**:207–210.

Steiner, C. C., A. S. Putnam, P. E. A. Hoeck, and O. A. Ryder 2013. Conservation genomics of threatened animal species. Annual Review of Animal Biosciences **1**:261–281.

Tewhey, R., V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork 2011. The importance of phase information for human genomics. Nature Reviews Genetics **12**:215–223.

The ENCODE Project Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489**:57–74.

The Heliconius Genome Consortium 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**:94–98.

Trapnell, C., and S. L. Salzberg 2009. How to map billions of short reads onto genomes. Nature Biotechnology **27**:455–457.

Tsai, I. J., T. D. Otto, and M. Berriman 2010. Method improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology **11**:R41.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics **43**:11.10.1–11.10.33.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild et al. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods **5**:247–252.

Vijay, N., J. W. Poelstra, A. Künstner, and J. B. W. Wolf 2013. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology **22**:620–634.

Vonk, F. J., N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. R. McCleary, H. M. E. Kerkkamp et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proceedings of the National Academy of Sciences **110**:20651–20656.

Wang, B., R. Ekblom, T. Strand, S. Portela-Bens, and J. Hoglund 2012. Sequencing of the core MHC region of black grouse (Tetrao tetrix) and comparative genomics of the galliform MHC. BMC Genomics **13**:553.

Wang, Z., J. Pascual-Anaya, A. Zadissa, W. Li, Y. Niimura, Z. Huang, C. Li et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nature Genetics **45**:701–706.

Wang, B., R. Ekblom, I. Bunikis, H. Siitari, and J. Hoglund 2014. Whole genome sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics **15**:180.

Warren, R. L., G. G. Sutton, S. J. M. Jones, and R. A. Holt 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics **23**:500–501.

Wei, F., Y. Hu, L. Zhu, M. W. Bruford, X. Zhan, and L. Zhang 2012. Black and white and read all over: the past, present and future of giant panda genetics. Molecular Ecology **21**:5660–5674.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature **452**:872–876.

Wolf, J. B. W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Molecular Ecology Resources **13**:559–572.

Wong, P., E. Wiley, W. Johnson, O. Ryder, S. O'Brien, D. Haussler, K.-P. Koepfli et al. 2012. Tissue sampling methods and standards for vertebrate genomics. GigaScience **1**:8.

Yandell, M., and D. Ence 2012. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics **13**:329–342.

Yim, H.-S., Y. S. Cho, X. Guang, S. G. Kang, J.-Y. Jeong, S.-S. Cha, H.-M. Oh et al. 2014. Minke whale genome and aquatic adaptation in cetaceans. Nature Genetics **46**:88–92.

Zavodna, M., C. E. Grueber, and N. J. Gemmell 2013. Parallel tagged next-generation sequencing on pooled samples – a new approach for population genetics in ecology and conservation. PLoS One **8**: e61471.

Zerbino, D. R., and E. Birney 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research **18**:821–829.

Zhan, X., S. Pan, J. Wang, A. Dixon, J. He, M. G. Muller, P. Ni et al. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nature Genetics **45**:563–566.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature **490**:49–54.

Zhou, X., and A. Rokas 2014. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. Molecular Ecology **23**:1679–1700.

Zimin, A., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke 2013. The MaSuRCA genome Assembler. Bioinformatics **29**:2669–2677.

 © 2014 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

D.6 Genomes

Further explanations for Chapter 6.

D.6.1 Background: genes, genomes and cells

- Clark, Douglas, Choi (OpenStax) (2018a) — Open textbook chapter on the double helix, base pairing, replication and DNA repair, at introductory undergraduate level. Useful for revising the nucleic-acid basics assumed by the rest of the chapter. *(book chapter; CC BY-NC-SA 4.0)*

  Chapter 14: DNA Structure and Function (Biology 2e). Clark, Douglas, Choi (OpenStax) (2018). *OpenStax Biology 2e.* https://openstax.org/books/biology-2e/pages/14-introduction

- Clark, Douglas, Choi (OpenStax) (2018b) — Open textbook chapter on the genetic code, transcription, RNA processing and translation in prokaryotes and eukaryotes. Covers the central dogma at the level summarised in this section. *(book chapter; CC BY-NC-SA 4.0)*

  Chapter 15: Genes and Proteins (Biology 2e). Clark, Douglas, Choi (OpenStax) (2018). *OpenStax Biology 2e.* https://openstax.org/books/biology-2e/pages/15-introduction

- Heather and Chain (2016) — A short historical review of DNA sequencing, from Sanger and Maxam-Gilbert sequencing to second- and third-generation methods. Gives background for the experimental methods mentioned in this section. *(review; CC BY 4.0)*

  The sequence of sequencers: The history of sequencing DNA. Heather & Chain (2016). *Genomics.* doi:10.1016/j.ygeno.2015.11.003

  > Determining the order of nucleic acid residues in biological samples is an integral component of a wide variety of research applications. Over the last fifty years large numbers of researchers have applied themselves to the production of techniques and technologies to facilitate this feat, sequencing DNA and RNA molecules. This time-scale has witnessed tremendous changes, moving from sequencing short oligonucleotides to millions of bases, from struggling towards the deduction of the coding sequence of a single gene to rapid and widely available whole genome sequencing. This article traverses those years, iterating through the different generations of sequencing technology, highlighting some of the key discoveries, researchers, and sequences along the way.

D.6.2 From the first genome projects

- Taylor et al. (2024) — Reviews how the incomplete reference from the Human Genome Project was completed by the Telomere-to-Telomere assembly and extended by the Human Pangenome Reference. Connects the genome projects in this section with the pangenome discussion in the next. *(review; CC BY 4.0)*

  Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Taylor et al. (2024). *Annual Review of Genomics and Human Genetics.* doi:10.1146/annurev-genom-021623-081639

  > The Human Genome Project was an enormous accomplishment, providing a foundation for countless explorations into the genetics and genomics of the human species. Yet for many years, the human genome reference sequence remained incomplete and lacked representation of human genetic diversity. Recently, two major advances have emerged to address these shortcomings: complete gap-free human genome sequences, such as the one developed by the Telomere-to-Telomere Consortium, and high-quality pangenomes, such as the one developed by the Human Pangenome Reference Consortium. Facilitated by advances in long-read DNA sequencing and genome assembly algorithms, complete human genome sequences resolve regions that have been historically difficult to sequence, including centromeres, telomeres, and segmental duplications. In parallel, pangenomes capture the extensive genetic diversity across populations worldwide. Together, these advances usher in a new era of genomics research,

enhancing the accuracy of genomic analysis, paving the path for precision medicine, and contributing to deeper insights into human biology.

- Birney (2021) — A brief perspective on why the Human Genome Project was organised internationally and kept its data freely available. Useful context for the Bermuda Principles and open data in genomics. *(perspective; CC BY 4.0)*

  The International Human Genome Project. Birney (2021). *Human Molecular Genetics*. doi:10.1093/hmg/ddab198

  > The human genome project was conceived and executed as an international project, due to both pragmatic and principled reasons. This internationality has served the project well, with the resulting human genome being freely available for all researchers in all countries. Over time the reference human genome will likely have to evolve to a graph genome, and tap into more diverse sequences worldwide. A similar international mindset underpins data analysis for the interpretation of the human genome from basic to clinical research.

- Pellicer et al. (2018) — Reviews the range of genome sizes in eukaryotes and especially land plants, and the roles of polyploidy and transposable elements in genome size change. Gives concrete examples for the C-value paradox. *(review; CC BY 4.0)*

  Genome Size Diversity and Its Impact on the Evolution of Land Plants. Pellicer et al. (2018). *Genes*. doi:10.3390/genes9020088

  > Genome size is a biodiversity trait that shows staggering diversity across eukaryotes, varying over 64,000-fold. Of all major taxonomic groups, land plants stand out due to their staggering genome size diversity, ranging ca. 2400-fold. As our understanding of the implications and significance of this remarkable genome size diversity in land plants grows, it is becoming increasingly evident that this trait plays not only an important role in shaping the evolution of plant genomes, but also in influencing plant community assemblages at the ecosystem level. Recent advances and improvements in novel sequencing technologies, as well as analytical tools, make it possible to gain critical insights into the genomic and epigenetic mechanisms underpinning genome size changes. In this review we provide an overview of our current understanding of genome size diversity across the different land plant groups, its implications on the biology of the genome and what future directions need to be addressed to fill key knowledge gaps.

D.6.3 Differences across the tree of life

- Youssef et al. (2019) — Open-access book chapter introducing genome structure and content across the tree of life, including how prokaryotic and eukaryotic genomes differ. A broad starting point for this section. *(book chapter; CC BY 4.0)*

  Introduction to Genome Biology and Diversity. Youssef et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_1

  > Organisms display astonishing levels of cell and molecular diversity, including genome size, shape, and architecture. In this chapter, we review how the genome can be viewed as both a structural and an informational unit of biological diversity and explicitly define our intended meaning of genetic information. A brief overview of the characteristic features of bacterial, archaeal, and eukaryotic cell types and viruses sets the stage for a review of the differences in organization, size, and packaging strategies of their genomes. We include a detailed review of genetic elements found outside the primary chromosomal structures, as these provide insights into how genomes are sometimes viewed as incomplete informational entities. Lastly, we reassess the definition of the genome in light of recent advancements in our understanding of the diversity of genomic structures and the mechanisms by which genetic information is expressed within the cell. Collectively, these topics comprise a good introduction to genome biology for the newcomer to the field and provide a valuable reference for those developing new statistical or computation methods in genomics. This review also prepares the reader for anticipated transformations in thinking as the field of genome biology progresses.

- Bobay and Ochman (2017) — Short review of why bacterial genomes are compact and gene-dense while eukaryotic genomes accumulate non-functional sequence, explained by a deletion bias in bacteria and by genetic drift. Complements the subsection on prokaryotic genome organisation and the population-genetic arguments in the next section. *(review; CC BY 4.0)*

  The Evolution of Bacterial Genome Architecture. Bobay & Ochman (2017). *Frontiers in Genetics*. doi:10.3389/fgene.2017.00072

  > The genome architecture of bacteria and eukaryotes evolves in opposite directions when subject to genetic drift, a difference that can be ascribed to the fact that bacteria exhibit a mutational bias that deletes superfluous sequences, whereas eukaryotes are biased toward large insertions. Expansion of eukaryotic genomes occurs through the addition of non-functional sequences, such as repetitive sequences and transposable elements, whereas variation in bacterial genome size is largely due to the acquisition and loss of functional accessory genes. These properties create the situation in which eukaryotes with very similar numbers of genes can have vastly different genome sizes, while in bacteria, gene number scales linearly with genome size. Some bacterial genomes, however, particularly those of species that undergo bottlenecks due to recent association with hosts, accumulate pseudogenes and mobile elements, conferring them a low gene content relative to their genome size. These non-functional sequences are gradually eroded and eliminated after long-term association with hosts, with the result that obligate symbionts have the smallest genomes of any cellular organism. The architecture of bacterial genomes is shaped by complex and diverse processes, but for most bacterial species, genome size is governed by a non-adaptive process, i.e., genetic drift coupled with a mutational bias toward deletions. Thus, bacteria with small effective population sizes typically have the smallest genomes. Some marine bacteria counter this near-universal trend: despite having immense population sizes, selection, not drift, acts to reduce genome size in response to metabolic constraints in their nutrient-limited environment.

- Koonin (2015) — Discusses the origin of the eukaryotic cell from an archaeal host and a bacterial endosymbiont, from a comparative-genomics perspective. Relevant to the subsection on organelle genomes and endosymbiosis. *(review; CC BY 4.0)*

  Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier?. Koonin (2015). *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi:10.1098/rstb.2014.0333

  > The origin of eukaryotes is a fundamental, forbidding evolutionary puzzle. Comparative genomic analysis clearly shows that the last eukaryotic common ancestor (LECA) possessed most of the signature complex features of modern eukaryotic cells, in particular the mitochondria, the endomembrane system including the nucleus, an advanced cytoskeleton and the ubiquitin network. Numerous duplications of ancestral genes, e.g. DNA polymerases, RNA polymerases and proteasome subunits, also can be traced back to the LECA. Thus, the LECA was not a primitive organism and its emergence must have resulted from extensive evolution towards cellular complexity. However, the scenario of eukaryogenesis, and in particular the relationship between endosymbiosis and the origin of eukaryotes, is far from being clear. Four recent developments provide new clues to the likely routes of eukaryogenesis. First, evolutionary reconstructions suggest complex ancestors for most of the major groups of archaea, with the subsequent evolution dominated by gene loss. Second, homologues of signature eukaryotic proteins, such as actin and tubulin that form the core of the cytoskeleton or the ubiquitin system, have been detected in diverse archaea. The discovery of this 'dispersed eukaryome' implies that the archaeal ancestor of eukaryotes was a complex cell that might have been capable of a primitive form of phagocytosis and thus conducive to endosymbiont capture. Third, phylogenomic analyses converge on the origin of most eukaryotic genes of archaeal descent from within the archaeal evolutionary tree, specifically, the TACK superphylum. Fourth, evidence has been presented that the origin of the major archaeal phyla involved massive acquisition of bacterial genes. Taken together, these findings make the symbiogenetic scenario for the origin of eukaryotes considerably more plausible and the origin of the organizational complexity of eukaryotic cells more readily explainable than they appeared until recently.

D.6.4 Genomes in populations

- Medini et al. (2020) — Opening chapter of an open-access book, written by the authors who introduced the pan-genome concept, describing how core and accessory genomes were defined

from multiple bacterial genomes. Other chapters of the same book cover pangenome ecology and selection. *(book chapter; CC BY 4.0)*

The Pangenome: A Data-Driven Discovery in Biology. Medini et al. (2020). *The Pangenome*. doi:10.1007/978-3-030-38281-0_1

> An early example of Big data in biology: how a mathematical model, developed to address a practical question in vaccinology, transformed established concepts, opening biology to the "unbounded."

- Wolf and Akey (2018) — Short review of what genome sequences have shown about Neanderthal and Denisovan admixture into modern humans, and which questions remain open. *(review; CC BY 4.0)*

  Outstanding questions in the study of archaic hominin admixture. Wolf & Akey (2018). *PLOS Genetics*. doi:10.1371/journal.pgen.1007349

  > The complete sequencing of archaic and modern human genomes has revolutionized the study of human history and evolution. The application of paleogenomics has answered questions that were beyond the scope of archaeology alone-definitively proving admixture between archaic and modern humans. Despite the remarkable progress made in the study of archaic-modern human admixture, many outstanding questions remain. Here, we review some of these questions, which include how frequent archaic-modern human admixture was in history, to what degree drift and selection are responsible for the loss and retention of introgressed sequences in modern human genomes, and how surviving archaic sequences affect human phenotypes.

- Ahlquist et al. (2021) — Reviews a decade of work on archaic introgression, including demographic history and the methods used to detect introgressed segments. Gives more methodological detail than the previous item. *(review; CC BY 4.0)*

  Our Tangled Family Tree: New Genomic Methods Offer Insight into the Legacy of Archaic Admixture. Ahlquist et al. (2021). *Genome Biology and Evolution*. doi:10.1093/gbe/evab115

  > The archaic ancestry present in the human genome has captured the imagination of both scientists and the wider public in recent years. This excitement is the result of new studies pushing the envelope of what we can learn from the archaic genetic information that has survived for over 50,000 years in the human genome. Here, we review the most recent ten years of literature on the topic of archaic introgression, including the current state of knowledge on Neanderthal and Denisovan introgression, as well as introgression from other as-yet unidentified archaic populations. We focus this review on four topics: 1) a reimagining of human demographic history, including evidence for multiple admixture events between modern humans, Neanderthals, Denisovans, and other archaic populations; 2) state-of-the-art methods for detecting archaic ancestry in population-level genomic data; 3) how these novel methods can detect archaic introgression in modern African populations; and 4) the functional consequences of archaic gene variants, including how those variants were co-opted into novel function in modern human populations. The goal of this review is to provide a simple-to-access reference for the relevant methods and novel data, which has changed our understanding of the relationship between our species and its siblings. This body of literature reveals the large degree to which the genetic legacy of these extinct hominins has been integrated into the human populations of today.

D.6.5 Repeats and the 3D genome

- Makałowski et al. (2019) — Open-access chapter describing the classes of transposable elements and the computational methods used to identify and annotate them. Useful for the subsection on transposable elements. *(book chapter; CC BY 4.0)*

  Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. Makałowski et al. (2019). *Evolutionary Genomics*. doi:10.1007/978-1-4939-9074-0_6

> Most genomes are populated by hundreds of thousands of sequences originated from mobile elements. On the one hand, these sequences present a real challenge in the process of genome analysis and annotation. On the other hand, they are very interesting biological subjects involved in many cellular processes. Here we present an overview of transposable elements biodiversity, and we discuss different approaches to transposable elements detection and analyses.

- Garrido-Ramos (2017) — Reviews the structure, organisation, evolution and possible functions of satellite DNA, including how sequencing data are used to study it. Complements the subsection on tandem repeats. *(review; CC BY 4.0)*

  Satellite DNA: An Evolving Topic. Garrido-Ramos (2017). *Genes*. doi:10.3390/genes8090230

  > Satellite DNA represents one of the most fascinating parts of the repetitive fraction of the eukaryotic genome. Since the discovery of highly repetitive tandem DNA in the 1960s, a lot of literature has extensively covered various topics related to the structure, organization, function, and evolution of such sequences. Today, with the advent of genomic tools, the study of satellite DNA has regained a great interest. Thus, Next-Generation Sequencing (NGS), together with high-throughput in silico analysis of the information contained in NGS reads, has revolutionized the analysis of the repetitive fraction of the eukaryotic genomes. The whole of the historical and current approaches to the topic gives us a broad view of the function and evolution of satellite DNA and its role in chromosomal evolution. Currently, we have extensive information on the molecular, chromosomal, biological, and population factors that affect the evolutionary fate of satellite DNA, knowledge that gives rise to a series of hypotheses that get on well with each other about the origin, spreading, and evolution of satellite DNA. In this paper, I review these hypotheses from a methodological, conceptual, and historical perspective and frame them in the context of chromosomal organization and evolution.

- Pal et al. (2018) — Explains how Hi-C data are generated and analysed, from normalisation to the calling of compartments, TADs and loops. Useful for understanding the methods behind 3D genome organisation. *(review; CC BY 4.0)*

  Hi-C analysis: from data generation to integration. Pal et al. (2018). *Biophysical Reviews*. doi:10.1007/s12551-018-0489-1

  > In the epigenetics field, large-scale functional genomics datasets of ever-increasing size and complexity have been produced using experimental techniques based on high-throughput sequencing. In particular, the study of the 3D organization of chromatin has raised increasing interest, thanks to the development of advanced experimental techniques. In this context, Hi-C has been widely adopted as a high-throughput method to measure pairwise contacts between virtually any pair of genomic loci, thus yielding unprecedented challenges for analyzing and handling the resulting complex datasets. In this review, we focus on the increasing complexity of available Hi-C datasets, which parallels the adoption of novel protocol variants. We also review the complexity of the multiple data analysis steps required to preprocess Hi-C sequencing reads and extract biologically meaningful information. Finally, we discuss solutions for handling and visualizing such large genomics datasets.

D.6.6 RNA genes and regulatory elements

- Palazzo and Gregory (2014) — A readable argument, based on genome size variation, transposable elements and population genetics, for why much of the human genome is likely to be non-functional. Directly relevant to the debate after ENCODE discussed in this section. *(review; CC BY 4.0)*

  The Case for Junk DNA. Palazzo & Gregory (2014). *PLoS Genetics*. doi:10.1371/journal.pgen.1004351

  > With the advent of deep sequencing technologies and the ability to analyze whole genome sequences and transcriptomes, there has been a growing interest in exploring putative functions of the very large fraction of the genome that is commonly referred to as "junk DNA."Whereas this is an issue of considerable importance in genome biology, there is an unfortunate tendency for researchers and science writers to proclaim the demise of junk DNA on a regular basis without properly addressing some of the fundamental issues that first led to the rise of the concept.In this review, we provide an overview of the major arguments

that have been presented in support of the notion that a large portion of most eukaryotic genomes lacks an organism-level function.Some of these are based on observations or basic genetic principles that are decades old, whereas others stem from new knowledge regarding molecular processes such as transcription and gene regulation.

- Ponting and Haerty (2022) — A critical review of the evidence for lncRNA function, discussing the pitfalls of experimental and computational methods. Argues that most lncRNA transcripts are likely to be noise or to have minor regulatory roles. *(review; CC BY 4.0)*

  Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. Ponting & Haerty (2022). *Annual Review of Genomics and Human Genetics*. doi:10.1146/annurev-genom-112921-123710

  > Do long noncoding RNAs (lncRNAs) contribute little or substantively to human biology? To address how lncRNA loci and their transcripts, structures, interactions, and functions contribute to human traits and disease, we adopt a genome-wide perspective. We intend to provoke alternative interpretation of questionable evidence and thorough inquiry into unsubstantiated claims. We discuss pitfalls of lncRNA experimental and computational methods as well as opposing interpretations of their results. The majority of evidence, we argue, indicates that most lncRNA transcript models reflect transcriptional noise or provide minor regulatory roles, leaving relatively few human lncRNAs that contribute centrally to human development, physiology, or behavior. These important few tend to be spliced and better conserved but lack a simple syntax relating sequence to structure and mechanism, and so resist simple categorization. This genome-wide view should help investigators prioritize individual lncRNAs based on their likely contribution to human biology.

- Marchese et al. (2017) — Reviews the proposed mechanisms by which lncRNAs act, through RNA sequence and structure, protein interactions and the act of transcription itself. A useful counterpoint to the more sceptical review above. *(review; CC BY 4.0)*

  The multidimensional mechanisms of long noncoding RNA function. Marchese et al. (2017). *Genome Biology*. doi:10.1186/s13059-017-1348-2

  > A major shift in our understanding of genome regulation has emerged recently. It is now apparent that the majority of cellular transcripts do not code for proteins, and many of them are long noncoding RNAs (lncRNAs). Increasingly, studies suggest that lncRNAs regulate gene expression through diverse mechanisms. We review emerging mechanistic views of lncRNAs in gene regulation in the cell nucleus. We discuss the functional interactions that lncRNAs establish with other molecules as well as the relationship between lncRNA transcription and function. While some of these mechanisms are specific to lncRNAs, others might be shared with other types of genes.

D.6.7 Gene order, synteny and genome rearrangements

- Liu et al. (2018) — Compares synteny-detection tools on real and simulated assemblies, and shows how assembly fragmentation and parameter choices change the inferred synteny blocks; a practical complement to the section. *(primary study; CC BY 4.0)*

  Inferring synteny between genome assemblies: a systematic evaluation. Liu et al. (2018). *BMC Bioinformatics*. doi:10.1186/s12859-018-2026-4

  > Background Genome assemblies across all domains of life are being produced routinely. Initial analysis of a new genome usually includes annotation and comparative genomics. Synteny provides a framework in which conservation of homologous genes and gene order is identified between genomes of different species. The availability of human and mouse genomes paved the way for algorithm development in large-scale synteny mapping, which eventually became an integral part of comparative genomics. Synteny analysis is regularly performed on assembled sequences that are fragmented, neglecting the fact that most methods were developed using complete genomes. It is unknown to what extent draft assemblies lead to errors in such analysis. Results We fragmented genome assemblies of model nematodes to various extents and conducted synteny identification and downstream analysis. We first show that synteny between species can be underestimated up to 40% and find disagreements between popular tools that infer synteny blocks. This inconsistency and further demonstration of erroneous gene ontology enrichment tests raise

questions about the robustness of previous synteny analysis when gold standard genome sequences remain limited. In addition, assembly scaffolding using a reference guided approach with a closely related species may result in chimeric scaffolds with inflated assembly metrics if a true evolutionary relationship was overlooked. Annotation quality, however, has minimal effect on synteny if the assembled genome is highly contiguous. Conclusions Our results show that a minimum N50 of 1 Mb is required for robust downstream synteny analysis, which emphasizes the importance of gold standard genomes to the science community, and should be achieved given the current progress in sequencing technology.

D.6.8 Pseudogenes, new genes and small proteins

- Van Oss and Carvunis (2019) — A short, clear review of how genes arise from non-coding DNA, proto-genes, and the methodological pitfalls in identifying de novo genes. *(review; CC BY 4.0)*

  De novo gene birth. Van Oss & Carvunis (2019). *PLOS Genetics*. doi:10.1371/journal.pgen.1008160

- Y.-W. Wang et al. (2020) — A worked example that weighs the different routes to new genes in one lineage, showing how each is detected. *(primary study; CC BY 4.0)*

  De Novo Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis inAmanita. Wang et al. (2020). *Genome Biology and Evolution*. doi:10.1093/gbe/evaa193

  By introducing novel capacities and functions, new genes and gene families may play a crucial role in ecological transitions. Mechanisms generating new gene families include de novo gene birth, horizontal gene transfer, and neofunctionalization following a duplication event. The ectomycorrhizal (ECM) symbiosis is a ubiquitous mutualism and the association has evolved repeatedly and independently many times among the fungi, but the evolutionary dynamics enabling its emergence remain elusive. We developed a phylogenetic workflow to first understand if gene families unique to ECM Amanita fungi and absent from closely related asymbiotic species are functionally relevant to the symbiosis, and then to systematically infer their origins. We identified 109 gene families unique to ECM Amanita species. Genes belonging to unique gene families are under strong purifying selection and are upregulated during symbiosis, compared with genes of conserved or orphan gene families. The origins of seven of the unique gene families are strongly supported as either de novo gene birth (two gene families), horizontal gene transfer (four), or gene duplication (one). An additional 34 families appear new because of their selective retention within symbiotic species. Among the 109 unique gene families, the most upregulated gene in symbiotic cultures encodes a 1-aminocyclopropane-1-carboxylate deaminase, an enzyme capable of downregulating the synthesis of the plant hormone ethylene, a common negative regulator of plant-microbial mutualisms.

The following open-access paper(s) are reproduced under their original licences.

OPEN ACCESS Freely available online

PLOS | GENETICS

Viewpoints

The Case for Junk DNA

Alexander F. Palazzo[1]*, T. Ryan Gregory[2]*

1 University of Toronto, Department of Biochemistry, Toronto, Ontario, Canada, **2** University of Guelph, Department of Integrative Biology, Guelph, Ontario, Canada

Overview

With the advent of deep sequencing technologies and the ability to analyze whole genome sequences and transcriptomes, there has been a growing interest in exploring putative functions of the very large fraction of the genome that is commonly referred to as "junk DNA." Whereas this is an issue of considerable importance in genome biology, there is an unfortunate tendency for researchers and science writers to proclaim the demise of junk DNA on a regular basis without properly addressing some of the fundamental issues that first led to the rise of the concept. In this review, we provide an overview of the major arguments that have been presented in support of the notion that a large portion of most eukaryotic genomes lacks an organism-level function. Some of these are based on observations or basic genetic principles that are decades old, whereas others stem from new knowledge regarding molecular processes such as transcription and gene regulation.

Introduction

The search for function in the genome

It has been known for several decades that only a small fraction of the human genome is made up of protein-coding sequences and that at least some noncoding DNA has important biological functions. In addition to coding exons, the genome contains sequences that are transcribed into functional RNA molecules (e.g., tRNA, rRNA, and snRNA), regulatory regions that control gene expression (e.g., promoters, silencers, and enhancers), origins of replication, and repeats that play structural roles at the chromosomal level (e.g., telomeres and centromeres).

New discoveries regarding potentially important sequences amongst the nonprotein-coding majority of the genome are becoming more prevalent. By far the best-known effort to identify functional regions in the human genome is the recently completed Encyclopaedia of DNA Elements (ENCODE) project [1], whose authors made the remarkable claim that a "biochemical function" could be assigned to 80% of the human genome [2]. Reports that ENCODE had refuted the existence of large amounts of junk DNA in the human genome received considerable media attention [3,4]. Criticisms that these claims were based on an extremely loose definition of "function" soon followed [5–8] (for a discussion of the relevant function concepts, see [9]), and debate continues regarding the most appropriate interpretation of the ENCODE results. Nevertheless, the excitement and subsequent backlash served to illustrate the widespread interest among scientists and nonspecialists in determining how much of the human genome is functionally significant at the organism level.

The origin of "junk DNA"

Although the term "junk DNA" was already in use as early as the 1960s [10–12], the term's origin is usually attributed to Susumu Ohno [13]. As Ohno pointed out, gene duplication can alleviate the constraint imposed by natural selection on changes to important gene regions by allowing one copy to maintain the original function as the other undergoes mutation. Rarely, these mutations will turn out to be beneficial, and a new gene may arise ("neofunctionalization") [14]. Most of the time, however, one copy sustains a mutation that eliminates its ability to encode a functional protein, turning it into a pseudogene. These sequences are what Ohno initially referred to as "junk" [13], although the term was quickly extended to include many types of noncoding DNA [15]. Today, "junk DNA" is often used in the broad sense of referring to any DNA sequence that does not play a functional role in development, physiology, or some other organism-level capacity. This broader sense of the term is at the centre of most current debate about the quantity—or even the existence—of "junk DNA" in the genomes of humans and other organisms.

It has now become something of a cliché to begin both media stories and journal articles with the simplistic claim that most or all noncoding DNA was "long dismissed as useless junk." The implication, of course, is that current research is revealing function in much of the supposed junk that was unwisely ignored as biologically uninteresting by past investigators. Yet, it is simply not true that potential functions for noncoding DNA were ignored until recently. In fact, various early commenters considered the notion that large swaths of the genome were nonfunctional to be "repugnant" [10,16], and possible functions were discussed each time a new type of nonprotein-coding sequence was identified (including pseudogenes, transposable elements, satellite DNA, and introns; for a compilation of relevant literature, see [17]).

Importantly, the concept of junk DNA was not based on ignorance about genomes. On the contrary, the term reflected *known* details about genome size variability, the mechanism of gene duplication and mutational degradation, and population genetics theory. Moreover, each of these observations and theoretical considerations remains valid. In this review, we examine several lines of evidence—both empirical and conceptual—that support the notion that a substantial percentage of the DNA in many eukaryotic genomes lacks an organism-level function and that the junk DNA concept remains viable post-ENCODE.

Citation: Palazzo AF, Gregory TR (2014) The Case for Junk DNA. PLoS Genet 10(5): e1004351. doi:10.1371/journal.pgen.1004351

Editor: Joshua M. Akey, University of Washington, United States of America

Published May 8, 2014

Copyright: © 2014 Palazzo, Gregory. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the Canadian Institutes for Health Research (CIHR) to AFP and a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to TRG. The funders had no role in the preparation of the article.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: alex.palazzo@utoronto.ca (AFP); rgregory@uoguelph.ca (TRG)

Genome Size and ''The Onion Test''

There are several key points to be understood regarding genome size diversity among eukaryotes and its relationship to the concept of junk DNA. First, genome size varies enormously among species [18,19]: at least 7,000-fold among animals and 350-fold even within vertebrates. Second, genome size varies independently of intuitive notions of organism complexity or presumed number of protein-coding genes (Figure 1). For example, a human genome contains eight times more DNA than that of a pufferfish but is 40 times smaller than that of a lungfish. Third, organisms that have very large genomes are not few in number or outliers—for example, of the >200 salamander genomes analyzed thus far, all are between four and 35 times larger than the human genome [18]. Fourth, even closely related species with very similar biological properties and the same ploidy level can differ significantly in genome size.

These observations pose an important challenge to any claim that most eukaryotic DNA is functional at the organism level. This logic is perhaps best illustrated by invoking ''the onion test'' [20]. The domestic onion, *Allium cepa*, is a diploid plant (2n = 16) with a haploid genome size of roughly 16 billion base pairs (16 Gbp), or about five times larger than humans. Although any number of species with large genomes could be chosen for such a comparison, the onion test simply asks: if most eukaryotic DNA is functional at the organism level, be it for gene regulation, protection against mutations, maintenance of chromosome structure, or any other such role, then why does an onion require five times more of it than a human? Importantly, the comparison is not restricted to onions versus humans. It could as easily be between pufferfish and lungfish, which differ by ~350-fold, or members of the genus *Allium*, which have more than a 4-fold range in genome size that is not the result of polyploidy [21].

In summary, the notion that the majority of eukaryotic noncoding DNA is functional is very difficult to reconcile with the massive diversity in genome size observed among species, including among some closely related taxa. The onion test is merely a restatement of this issue, which has been well known to genome biologists for many decades [18].

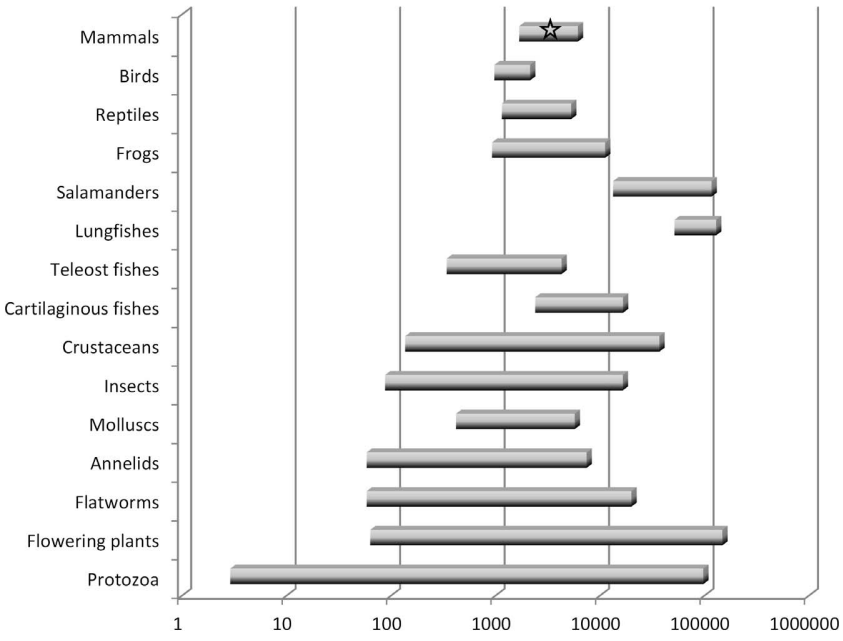


Figure 1. Summary of haploid nuclear DNA contents (''genome sizes'') for various groups of eukaryotes. This graph is based on data for about 10,000 species [18,19]. There is a wide range in genome sizes even among developmentally similar species, and there is no correspondence between genome size and general organism complexity. Humans, which have an average-sized genome for a mammal, are indicated by a star. Note the logarithmic scale.
doi:10.1371/journal.pgen.1004351.g001

Genome Composition

Another important consideration is the composition of eukaryotic genomes. Far from being composed of mysterious ''dark matter,'' the characteristics of the sequences constituting 98% or so of the human genome that is nonprotein-coding are generally well understood.

Transposable elements

By far the dominant type of nongenic DNA are transposable elements (TEs), including various well-described retroelements such as Short and Long Interspersed Nuclear Elements (SINEs and LINEs), endogenous retroviruses, and cut-and-paste DNA transposons. Because of their capacity to increase in copy number, transposable elements have long been described as ''parasitic'' or ''selfish'' [22,23]. However, the vast majority of these elements are inactive in humans, due to a very large fraction being highly degraded by mutation. Due to this degeneracy, estimates of the proportion of the human genome occupied by TEs has varied widely, between one-half and two-thirds [24,25]. Larger genomes, such as those of salamanders and lungfishes, almost certainly contain an even more enormous quantity of transposable element DNA [26,27].

Many examples have been found in which TEs have taken on regulatory or other functional roles in the genome [28]. In recognition of the more complex interactions between transposable elements and their hosts, Kidwell and Lisch proposed an expansion of the ''parasitism'' framework where each TE can be classified along a spectrum from parasitism to mutualism [29]. Nevertheless, there is evidence of organism-level function for only a tiny minority of TE sequences. It is therefore not obvious that functional explanations can be extrapolated from a small number of specific examples to all TEs within the genome.

Highly repetitive DNA

Another large fraction of the genome consists of highly repetitive DNA. These regions are extremely variable even amongst individuals of the same population (hence their use as ''DNA fingerprints'') and can expand or contract through processes such as unequal crossing over or replication slippage. Many repeats are thought to be derived from truncated TEs, but others consist of tandem arrays of di- and trinucleotides [30]. As with TEs, some highly repetitive sequences play a role in gene regulation

(for example, [31]). Others, such as telomeric- and centromeric-associated repeats [32,33], play critical roles in chromosomal maintenance. Despite this, there is currently no evidence that the majority of highly repetitive elements are functional.

Introns

According to Gencode v17, about 40% of the human genome is comprised of intronic regions; however, this figure is likely an overestimate as it includes all annotated events. It is also important to note that a large fraction of TEs and repetitive elements are found in introns. Although introns can increase the diversity of protein products by modulating alternative splicing, it is also clear that the vast majority of intronic sequence evolves in an unconstrained way, accumulating mutations at about the same rate as neutral regions. Although the median intron size in humans is ~1.5 kb [30], data suggest that most of the constrained sequence is confined to the first and last 150 nucleotides [34].

Pseudogenes

The human genome is also home to a large number of pseudogenes. Estimates of the total number range from 12,600 to 19,700 [35]. These include both "classical" pseudogenes (direct duplicates, of the sort imagined by Ohno [13]) and "processed" pseudogenes, which are reverse transcribed from mRNA [36]. Once again, although some pseudogenes have been co-opted for organism-level function (for example see [37]), most are simply evolving without selective constraints on their sequences and likely have no function [38].

Conserved sequences

Several analyses of sequence conservation between humans and other mammals have found that about 5% of the genome is conserved [1,39–42]. It is possible that an additional 4% of the human genome is under lineage-specific selection pressure [39]; however, this estimate appears to be somewhat questionable [43,44] (also see [45]). Ignoring these problems, the idea that 9% of the human genome shows signs of functionality is actually consistent with the results of ENCODE and other large-scale genome analyses.

Besides protein-coding sequences (including associated untranslated regions), which make up 1.5%–2.5% of the human genome [24], data from ENCODE suggest that conserved long noncoding RNAs (lncRNAs) are generated from about 9,000 loci that add up to less than an additional 0.4% [46,47]. Thus, even if a vast new untapped world of functional noncoding RNA is discovered, this will probably be transcribed from a small fraction of the human genome.

At first blush, sequences that are bound by transcription factors (TFs) appear to be very abundant, making up about 8.5% of the genome according to ENCODE [2]. This number, however, is an estimate of regions that are hypersensitive to DNase I treatment due to the displacement of nucleosomes by TFs. As pointed out by others [6], these regions are annotated as being several hundreds of nucleotides long and are thus much larger than the actual size of individual TF-binding motifs, which are typically 10 bp in length [48]. By ENCODE's own estimates, less than half of the nucleotide bases in these DNase I hypersensitivity regions contain actual TF recognition motifs [2], and only 60% of these are under purifying selection [49]. Others have found that weak and transient TF-binding events are routinely identified by chromatin IP experiments despite the fact that they do not significantly contribute to gene expression [50–53] and are poorly conserved [53]. Given that experiments performed in a diverse number of eukaryotic systems have found only a small correlation between TF-binding events and mRNA expression [54,51], it appears that in most cases only a fraction of TF-binding sites significantly impacts local gene expression.

In summary, most of the major constituents of the genome have been well characterized. The majority of human DNA consists of repetitive, mutationally degraded sequences. There are unambiguous examples of nonprotein-coding sequences of various types having been co-opted for organism-level functions in gene regulation, chromosome structure, and other roles, but at present evidence from the published literature suggests that these represent a small minority of the human genome.

Evolutionary Forces

To understand the current state of the human genome, we need to examine how it evolved, and as Michael Lynch once wrote, "Nothing in evolution makes sense except in the light of population genetics" [55]. Unfortunately, concepts that have been generated by this field have not been widely recognized in other domains of the life sciences. In particular, what is underappreciated by many nonevolution specialists is that much of molecular evolution in eukaryotes is primarily the result of genetic drift, or the fixation of neutral mutations. This view has been widely appreciated by molecular evolutionary biologists for the past 35 years.

The nearly neutral theory of molecular evolution

An important development in the understanding of how various evolutionary forces shape eukaryotic genes and genomes came with the theories developed by Kimura, Ohta, King, and Jukes. They demonstrated that alleles that were slightly beneficial or deleterious behaved like neutral alleles, provided that the absolute value of their selection coefficient was smaller than the inverse of the "effective" population size [56–59]. In other words, it is important to keep in mind population size when thinking about whether deleterious mutations are subjected to purifying selection.

It is also important to realize that the "effective" population size is dependent on many factors and is typically much lower than the total number of individuals in a species [55]. For humans it has been estimated that the historical effective population size is approximately 10,000, and this is on the low side in comparison to most metazoans [60]. Given the overall low figures for multicellular organisms in general, we would expect that natural selection would be powerless to stop the accumulation of certain genomic alterations over the entirety of metazoan evolution. One type of mutation that fits this description is intergenic insertions, be they transposable elements, pseudogenes, or random sequence [55]. The creation and loss of TF-binding motifs or cryptic transcriptional start sites in these same intergenic regions will equally be invisible to natural selection, provided that these do not drastically alter the expression of any nearby genes or cause the production of stable toxic transcripts. Thus, a central tenet of the nearly neutral theory of molecular evolution is that extraneous DNA sequences can be present within genomes, provided that they do not significantly impact the fitness of the organism.

Genetic load

It has long been appreciated that there is a limit to the number of deleterious mutations that an organism can sustain per generation [61,62]. The presence of these mutations is usually not harmful, because diploid organisms generally require only one functional copy of any given gene. However, if the rate at which these mutations are generated is higher than the rate at which natural selection can weed them out, then the collective

genomes of the organisms in the species will suffer a meltdown as the total number of deleterious alleles increases with each generation [63]. This rate is approximately one deleterious mutation per generation. In this context it becomes clear that the overall mutation rate would place an upper limit to the amount of functional DNA. Currently, the rate of mutation in humans is estimated to be anywhere from 70–150 mutations per generation [64,65]. By this line of reasoning, we would estimate that, at most, only 1% of the nucleotides in the genome are essential for viability in a strict sequence-specific way. However, more recent computational models have demonstrated that genomes could sustain multiple slightly deleterious mutations per generation [66]. Using statistical methods, it has been estimated that humans sustain 2.1–10 deleterious mutations per generation [66–68]. These data would suggest that at most 10% of the human genome exhibits detectable organism-level function and conversely that at least 90% of the genome consists of junk DNA. These figures agree with measurements of genome conservation (~9%, see above) and are incompatible with the view that 80% of the genome is functional in the sense implied by ENCODE. It remains possible that large amounts of noncoding DNA play structural or other roles independent of nucleotide sequence, but it far from obvious how this would be reconciled with "the onion test."

The evolution of the nucleus

When dealing with the evolution of any lineage, one must also keep in mind unique events, also known as historical contingencies, which constrain and shape subsequent evolutionary trajectories [69]. One of these key events in our own ancestry was the evolution of the eukaryotic nucleus. A further examination of why the nucleus evolved and how this altered cellular function may generate significant insights into the current shape of the eukaryotic genome.

One important event in early eukaryotic evolution was the development of a symbiotic relationship between the α-proteobacteria progenitor of mitochondria and an archaebacteria-like host [70,71]. As with most endosymbiotically derived organelles [72], DNA was transferred from mitochondria to the host. In this way, Group II introns, which are still found in both mitochondria and α-proteobacteria [73], invaded the host genome. Group II introns are parasitic DNA fragments that replicate when they are transcribed, typically as part of a larger transcript. The intron then folds into a catalytic ribozyme that splices itself out of the precursor transcript and then reinserts itself at a new genomic locus by reversing the splicing reaction. Importantly, functional fragments of Group II introns can splice out inactive versions in a trans-splicing reaction [74,75]. As described elsewhere, it is likely that Group II introns proliferated and evolved into two populations: inactivated copies that could be nonetheless spliced out in trans, and active fragments that promoted splicing of the former group. This latter group eventually evolved into the spliceosomal snRNAs [75–77]. This idea is supported by not only structural, catalytic, and functional similarities between Group II introns and snRNAs [78,79] but also by the fact that expression of the U5 snRNA rescues the splicing of Group II introns that lack the corresponding U5-like region [80].

It is likely that the proliferation of trans-splicing triggered the spatial segregation of RNA processing (the nucleoplasm) from the translation machinery (the cytoplasm) [77]. This subdivision ensured that mRNAs were properly spliced before they encountered the translation machinery. Not only would this segregation prevent translating ribosomes from interfering with the splicing reaction (and vice versa) but would also prevent the translation of incompletely processed mRNAs, which often encode toxic proteins [81,82]. Importantly, the segregation of translation from both transcription and RNA processing provided an opportunity for nuclear quality-control processes to eliminate misprocessed and spurious transcripts that did not meet the minimal requirements of "mRNA identity" (see below) before these RNAs ever encountered a ribosome. This in turn permitted intergenic DNA and cryptic transcriptional start sites to proliferate with minimal cost to the fitness of the organism. It should also be noted that the increase in ATP regeneration due to mitochondrial-derived metabolic pathways provided the surplus energy that is required to support an expansion not only in genome size and membranes [83,84] but also wasteful transcription. Thus, by several independent mechanisms, the acquisition of mitochondria likely allowed the expansion of nonfunctional intergenic DNA and the evolution of a noisy transcriptional system.

Gene Expression in Eukaryotes

Eukaryotic transcription is inherently noisy

One of the most widely discussed discoveries of the past decade of transcriptome analysis is that much of the metazoan genome is transcribed at some level (although this, too, was already recognized in rough outline in the 1970s [15]). When nascent transcripts from mouse have been analyzed by deep sequencing, the total number of reads that map to intergenic loci is almost equivalent to the number mapping to exonic regions (Figure 2A, reproduced from reference [85]). This is consistent with the observation that a large fraction of the cellular pool of RNA Polymerase II is associated with intergenic regions [86] and that transcription can be initiated at random sequences (see Figure S4 in [87]) and nucleosome-free regions [88,89]. Strikingly, when one examines the steady state level of polyadenylated RNA, very little maps to intergenic regions (Figure 2A, 2B, the latter reproduced from reference [46]; also see [85,90–92]). In fact, when one eliminates the ~9,000 transcript species that are thought to be derived from conserved lncRNA, then most of the annotated noncoding polyadenylated RNAs are present at levels below one copy per cell and are found exclusively in the nucleus (Figure 2B). The situation is no better in the unpolyadenylated pool, in which the amount of lncRNA and intergenic RNA is practically insignificant, especially in the cytoplasmic pool (Figure 2B). In aggregate, these data indicate that the majority of intergenic RNAs are degraded almost immediately after transcription. Consistent with this idea, the level of intergenic transcripts increase when RNA degradation machinery is inhibited [93–101]. Although pervasive transcription has been used as an argument against junk DNA [3,4], it is in fact entirely in line with the idea that intergenic regions are evolving under little-to-no constraint, especially when one considers that this intergenic transcription is unstable.

Identifying mRNA from intergenic transcription

A common theme that has emerged from the study of mRNA synthesis is that various steps in RNA synthesis and processing are biochemically coupled. In other words, cellular machineries that participate in one biochemical activity also promote subsequent steps. For example, during the splicing of the 5′most intron, the spliceosome collaborates with the 5′cap binding complex to deposit nuclear export factors onto the 5′end of the processed transcript [102,103], and this helps to explain why splicing enhances the nuclear export of mRNA [104–106]. Countless other examples of coupling exist (for reviews, see [107–111]).

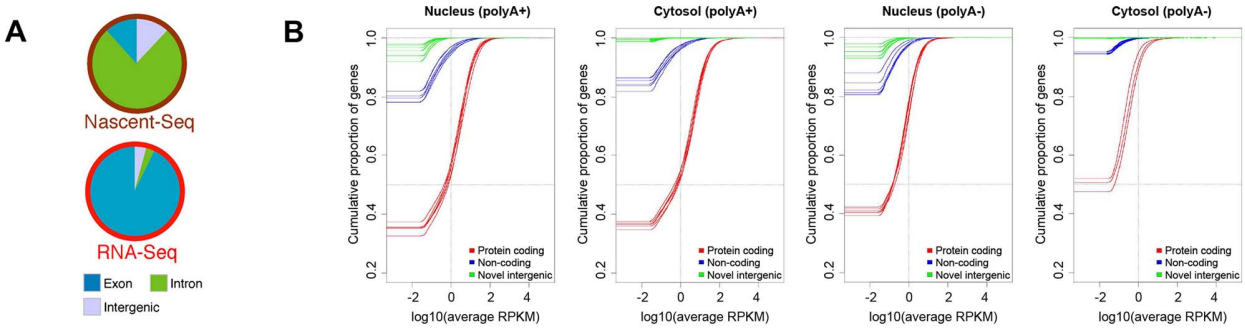


Figure 2. Levels of protein-coding and intergenic RNAs in mammalian cells. (A) Analysis of nascent and total poly(A)+ RNA levels from mouse liver nuclei. Nascent (i.e., polymerase-associated) RNA and poly(A)+ RNA were isolated from mouse liver nuclei and analyzed by high-throughput sequencing. Individual reads were categorized by their source. Exonic and intronic are from known referenced genes (i.e., "RefSeq" genes), while intergenic originate from nonreferenced loci (i.e., "non-RefSeq") in the mouse genome. Reproduced from [85]. (B) Empirical Cumulative Distribution Function (ECDF) of transcript expression in each cell compartment as determined by the ENCODE consortia. Results for RNA that either contain ("polyA+") or lack ("polyA−") a poly(A)-tail in the nucleus and cytosolic fractions are shown. Each human cell line that was analyzed is represented by three lines, one for each pool of RNA (red for protein-coding RNAs, blue for lncRNAs ["noncoding"], and green for intergenic transcripts ["novel intergenic"]). The lines indicate the cumulative fraction of RNAs in a given pool (y-axis) that are expressed at levels that are equal or less than the reads per kilobase per million mapped reads (RPKM) on the x-axis. Total numbers in each pool are as follows: reference protein coding genes: 20,679, loci producing lncRNAs: 9,277, and regions producing intergenic transcripts: 41,204. Transcripts with expression levels of 0 RPKM were adjusted to an artificial value of 10^{-6} RPKM so that the onset of each graph represents the fraction of nonexpressed genes or loci. Note that 1–4 RPKM is approximately equivalent to one copy per tissue culture cell [46,129]. Using this figure, one can easily deduce that the vast majority of intergenic transcripts are present at levels less than one copy per cell. Reproduced with permission from [46].
doi:10.1371/journal.pgen.1004351.g002

The ultimate goal of these coupling reactions is to sort protein-coding RNAs (i.e. mRNA) from intergenic transcripts [111,112]. Given that, on average, protein-coding genes have eight introns [30], while the majority of annotated ENCODE intergenic transcripts tend not to be spliced [46], introns help distinguish these two populations and thus serve as "mRNA identity" markers. These mRNA identity features activate coupling reactions, which in turn promote the further processing, nuclear export, and translation of a particular transcript. Likewise, other classes of functional RNAs (e.g., tRNAs and snRNAs) have their own identity elements [113]. In contrast, transcripts that lack identity elements are targeted for degradation. In agreement with this model, intronless RNA molecules that have a random sequence are poorly exported from the nucleus and have a very short half-life [114,115]. In contrast, intronless mRNAs have specialized motifs that promote their nuclear export [105,116–119].

In light of the fact that many functional lncRNAs serve a role in regulating chromatin structure or transcription, it is not surprising that most localise to the nucleoplasm [46]. One would predict that lncRNAs contain a differential set of identity elements that not only serve to prevent their decay but also retain them in the nucleus. This would especially be critical for lncRNAs that are spliced. Despite this, the elements that regulate the localization and stability of these RNAs have received little attention, but can be informed by the view that they may have their own identity markers.

It is also important to point out that eukaryotes have other mechanisms that either degrade aberrant mRNAs (e.g., nonsense-mediated decay) or limit the amount of intergenic transcription (e.g., heterochromatin). Nevertheless, eukaryotes appear to have evolved an intricate network of coupling reactions that are required to cope with a large burden of junk RNA. These findings are consistent with the idea that eukaryotic genomes are filled with junk DNA that is transcribed at a low level.

An alternative view of transcription and conservation?

In an attempt to counter the argument that sequence conservation is a prerequisite for functionality, it has been recently proposed that certain transcriptional events may serve some role in regulating cellular function, despite the fact that the sequence of the transcriptional product is unconstrained [120]. Indeed, this view is in line with the findings that the transcription of certain yeast genes is inhibited as a consequence of the production of cryptic unstable transcripts originating from upstream and/or downstream promoters (for a review see [121]). Other examples have linked the generation of cryptic unstable transcripts to chromatin modifications [101,122], DNA methylation [123], and DNA stability [124]. However, it remains unclear whether the majority of unstable noncoding RNAs have any effect on DNA or chromatin, let alone contribute to the fitness of the organism. In the cases where cryptic unstable transcriptional events impact gene expression, they usually consist of short transcripts that are synthesized from regions around the transcriptional start sites or within the gene itself [121]. Indeed most of the available data are consistent with the fact that transcriptional start sites are promiscuous, often generating bidirectional transcription [100,101], and that subsequent coupling processes, such as the interaction between promoter-associated complexes and 3′end processing factors, are required to enforce proper transcriptional directionality [125]. Other unstable transcripts function to promote or maintain heterochromatin formation in the vicinity of the transcriptional site, likely because these regions produce toxic transcripts [122]. Although this form of transcription has a function (viz., to maintain a repressive state), it is not clear that the elimination of these regions would have any effect on the organism [8]. The transcription of other short unstable transcripts, mostly produced from enhancer regions, has been shown to promote gene expression [126]; however, again these "enhancer RNAs" are transcribed from a small fraction of the total genome [127]. As stated by others [128], it is imperative that those who claim that the vast majority of intergenic transcription is functional test their hypotheses. In the absence of this evidence, the

declaration that we are in the midst of a paradigm shift with regards to eukaryotic genomes and gene expression [120] seems premature.

Concluding Remarks

For decades, there has been considerable interest in determining what role, if any, the majority of the DNA in eukaryotic genomes plays in organismal development and physiology. The ENCODE data are only the most recent contribution to a long-standing research program that has sought to address this issue. However, evidence casting doubt that most of the human genome possesses a functional role has existed for some time. This is not to say that none of the nonprotein-coding majority of the genome is functional—examples of functional noncoding sequences have been known for more than half a century, and even the earliest proponents of "junk DNA" and "selfish DNA" predicted that further examples would be found. Nevertheless, they also pointed out that evolutionary considerations, information regarding genome size diversity, and knowledge about the origins and features of genomic components do not support the notion that all of the DNA must have a function by virtue of its mere existence. Nothing in the recent research or commentary on the subject has challenged these observations.

Acknowledgments

We would like to thank L. Moran, S. Eddy, D. Graur, R. Hardison, J. Wan, and A. Akef for helpful comments on the manuscript.

References

1. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816. doi:10.1038/nature05874
2. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. doi:10.1038/nature11247
3. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, et al. (2012) Genomics: ENCODE explained. Nature 489: 52–55. doi:10.1038/489052a
4. Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. Science 337: 1159, 1161. doi:10.1126/science.337.6099.1159
5. Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. Curr Biol CB 22: R898–899. doi:10.1016/j.cub.2012.10.002
6. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, et al. (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5: 578–590. doi:10.1093/gbe/evt028
7. Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A 110: 5294–5300. doi:10.1073/pnas.1221376110
8. Niu D-K, Jiang L (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun 430: 1340–1343. doi:10.1016/j.bbrc.2012.12.074
9. Elliott TA, Linquist S, Gregory TR (2014) Conceptual and empirical challenges of ascribing functions to transposable elements. Am Nat In press.
10. Aronson AI, Bolton ET, Britten RJ, Cowie DB, Duerksen JD, et al. (1960) Biophysics. Year book - Carnegie Institution of Washington (1960). Volume 59. Baltimore, MD: Lord Baltimore Press. pp. 229–289.
11. Ehert CF, de Haller G (1963) Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium. J Ultrastruct Res 23: SUPPL6: 1–42.
12. Graur D (2013) The Origin of Junk DNA: A Historical Whodunnit. Judge Starling. Available: http://judgestarling.tumblr.com/post/64504735261/the-origin-of-junk-dna-a-historical-whodunnit. Accessed 23 December 2013.
13. Ohno S (1972) So much "junk" DNA in our genome. In: Smith HH, editor. Evolution of Genetic Systems. New York: Gordon and Breach. pp. 366–370.
14. Ohno S (1970) Evolution by gene duplication. London, New York: Allen & Unwin; Springer-Verlag. 160 p.
15. Comings DE (1972) The structure and function of chromatin. Adv Hum Genet 3: 237–431.
16. Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Science 161: 529–540.
17. Gregory TR (2008) Junk DNA – the quotes of interest series. Available: http://www.genomicron.evolverzone.com/2008/02/junk-dna-quotes-of-interest-series/. Accessed 10 April 2014.
18. Gregory TR (2013) Animal Genome Size Database. Available: http://www.genomesize.com. Accessed 10 April 2014.
19. Bennett MD, Leitch IJ (2012) Plant DNA C-values Database (Release 6.0, Dec. 2012). Available: http://data.kew.org/cvalues/. Accessed 10 April 2014.
20. Gregory TR (2007) The onion test. Available: http://www.genomicron.evolverzone.com/2007/04/onion-test/. Accessed 10 April 2014.
21. Ricroch A, Yockteng R, Brown SC, Nadot S (2005) Evolution of genome size across some cultivated Allium species. Genome Natl Res Counc Can Génome Cons Natl Rech Can 48: 511–520. doi:10.1139/g05-017
22. Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284: 604–607.
23. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601–603.
24. Gregory TR (2005) Synergy between sequence and size in large-scale genomics. Nat Rev Genet 6: 699–708. doi:10.1038/nrg1674
25. De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7: e1002384. doi:10.1371/journal.pgen.1002384
26. Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, et al. (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol Evol 4: 168–183. doi:10.1093/gbe/evr139
27. Metcalfe CJ, Filée J, Germon I, Joss J, Casane D (2012) Evolution of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1 and L2 LINE elements. Mol Biol Evol 29: 3529–3539. doi:10.1093/molbev/mss159
28. Cowley M, Oakey RJ (2013) Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet 9: e1003234. doi:10.1371/journal.pgen.1003234
29. Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evol Int J Org Evol 55: 1–24.
30. Scherer S (2008) A short guide to the human genome. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press. 173 p.
31. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42: 631–634. doi:10.1038/ng.600
32. Hemann MT, Strong MA, Hao LY, Greider CW (2001) The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. Cell 107: 67–77.
33. Torras-Llort M, Moreno-Moreno O, Azorín F (2009) Focus on the centre: the role of chromatin on the regulation of centromere identity and function. EMBO J 28: 2337–2348. doi:10.1038/emboj.2009.174
34. Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A (2007) Patterns and rates of intron divergence between humans and chimpanzees. Genome Biol 8: R21. doi:10.1186/gb-2007-8-2-r21
35. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. (2012) The GENCODE pseudogene resource. Genome Biol 13: R51. doi:10.1186/gb-2012-13-9-r51
36. Zhang Z, Gerstein M (2004) Large-scale analysis of pseudogenes in the human genome. Curr Opin Genet Dev 14: 328–335. doi:10.1016/j.gde.2004.06.003
37. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell 146: 353–358. doi:10.1016/j.cell.2011.07.014
38. Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 23: 219–224. doi:10.1016/j.tig.2007.03.003
39. Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675–1678. doi:10.1126/science.1225057
40. Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? Genome Res 21: 1769–1776. doi:10.1101/gr.116814.110
41. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482. doi:10.1038/nature10530
42. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901–913. doi:10.1101/gr.3577405
43. Bray N, Pachter L (2012) Comment on "Evidence of Abundant and Purifying Selection in Humans for Recently Acquired Regulatory Functions". Cornell University Library. arXiv:1212.3076 [q-bio.GN]. Available: http://arxiv.org/abs/1212.3076. Accessed 10 April 2014.
44. Green P, Ewing B (2013) Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions." Science 340: 682. doi:10.1126/science.1233195
45. Ward LD, Kellis M (2013) Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions." Science 340: 682. doi:10.1126/science.1233366
46. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi:10.1038/nature11233
47. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. (2012) The GENCODE v7

catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22: 1775–1789. doi:10.1101/gr.132159.111
48. Stewart AJ, Hannenhalli S, Plotkin JB (2012) Why transcription factor binding sites are ten nucleotides long. Genetics 192: 973–985. doi:10.1534/genetics.112.143370
49. Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, et al. (2012) Personal and population genomics of human regulatory variation. Genome Res 22: 1689–1697. doi:10.1101/gr.134890.111
50. Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484: 251–255. doi:10.1038/nature10985
51. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. Dev Cell 21: 611–626. doi:10.1016/j.devcel.2011.09.008
52. Li X, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6: e27. doi:10.1371/journal.pbio.0060027
53. Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, et al. (2013) Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. PLoS Genet 9: e1003748. doi:10.1371/journal.pgen.1003748
54. Spitz F, Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13: 613–626. doi:10.1038/nrg3207
55. Lynch M (2007) The origins of genome architecture. Sunderland Mass.: Sinauer Associates. 494 p.
56. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
57. King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164: 788–798.
58. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.
59. Kimura M (1984) The Neutral theory of molecular evolution. Cambridge [Cambridgeshire]; New York: Cambridge University Press. 367 p.
60. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205. doi:10.1038/nrg2526
61. Muller HJ (1950) Our load of mutations. Am J Hum Genet 2: 111–176.
62. Knudson AG Jr (1979) Presidential address. Our load of mutations and its burden of disease. Am J Hum Genet 31: 401–413.
63. Lynch M, Conery J, Burger R (1995) Mutational meltdowns in sexual populations. Evolution 49: 1067–1080.
64. Keightley PD (2012) Rates and fitness consequences of new mutations in humans. Genetics 190: 295–304. doi:10.1534/genetics.111.134668
65. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13: 745–753. doi:10.1038/nrg3295
66. Lesecque Y, Keightley PD, Eyre-Walker A (2012) A resolution of the mutation load paradox in humans. Genetics 191: 1321–1330. doi:10.1534/genetics.112.140343
67. Eory L, Halligan DL, Keightley PD (2010) Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol 27: 177–192. doi:10.1093/molbev/msp219
68. Reed FA, Akey JM, Aquadro CF (2005) Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res 15: 1211–1221. doi:10.1101/gr.3413205
69. Gould SJ (1994) The evolution of life on the earth. Sci Am 271: 84–91.
70. Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14: 255–274.
71. Woese CR (1977) Endosymbionts and mitochondrial origins. J Mol Evol 10: 93–96.
72. Martin W (2003) Gene transfer from organelles to the nucleus: frequent and in big chunks. Proc Natl Acad Sci U S A 100: 8612–8614. doi:10.1073/pnas.1633606100
73. Ferat JL, Michel F (1993) Group II self-splicing introns in bacteria. Nature 364: 358–361. doi:10.1038/364358a0
74. Jarrell KA, Dietrich RC, Perlman PS (1988) Group II intron domain 5 facilitates a trans-splicing reaction. Mol Cell Biol 8: 2361–2366.
75. Stoltzfus A (1999) On the possibility of constructive neutral evolution. J Mol Evol 49: 169–181.
76. Hickey DA, Benkel BF, Abukashawa SM (1989) A general model for the evolution of nuclear pre-mRNA introns. J Theor Biol 137: 41–53.
77. Martin W, Koonin EV (2006) Introns and the origin of nucleus-cytosol compartmentalization. Nature 440: 41–45. doi:10.1038/nature04531
78. Toor N, Keating KS, Taylor SD, Pyle AM (2008) Crystal structure of a self-spliced group II intron. Science 320: 77–82. doi:10.1126/science.1153803
79. Keating KS, Toor N, Perlman PS, Pyle AM (2010) A structural analysis of the group II intron active site and implications for the spliceosome. RNA 16: 1–9. doi:10.1261/rna.1791310
80. Hetzer M, Wurzer G, Schweyen RJ, Mueller MW (1997) Trans-activation of group II intron splicing by nuclear U5 snRNA. Nature 386: 417–420. doi:10.1038/386417a0
81. Cali BM, Anderson P (1998) mRNA surveillance mitigates genetic dominance in Caenorhabditis elegans. Mol Gen Genet 260: 176–184.
82. Khajavi M, Inoue K, Lupski JR (2006) Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. Eur J Hum Genet 14: 1074–1081. doi:10.1038/sj.ejhg.5201649
83. Lane N, Martin W (2010) The energetics of genome complexity. Nature 467: 929–934. doi:10.1038/nature09486
84. Lane N (2011) Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct 6: 35. doi:10.1186/1745-6150-6-35
85. Menet JS, Rodriguez J, Abruzzi KC, Rosbash M (2012) Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. eLife 1: e00011. doi:10.7554/eLife.00011.
86. Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14: 103–105. doi:10.1038/nsmb0207-103
87. White MA, Myers CA, Corbo JC, Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci U S A 110: 11952–11957. doi:10.1073/pnas.1307449110
88. Cheung V, Chua G, Batada NN, Landry CR, Michnick SW, et al. (2008) Chromatin- and transcription-related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome. PLoS Biol 6: e277. doi:10.1371/journal.pbio.0060277
89. Buratowski S (2008) Transcription. Gene expression–where to start? Science 322: 1804–1805. doi:10.1126/science.1168805
90. Babak T, Blencowe BJ, Hughes TR (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. BMC Genomics 6: 104. doi:10.1186/1471-2164-6-104
91. Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 5: e1000598. doi:10.1371/journal.pcbi.1000598
92. Van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol 8: e1000371. doi:10.1371/journal.pbio.1000371
93. Wyers F, Rougemaille M, Badis G, Rousselle J-C, Dufour M-E, et al. (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121: 725–737. doi:10.1016/j.cell.2005.04.030
94. Davis CA, Ares M Jr (2006) Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 103: 3262–3267. doi:10.1073/pnas.0507783103
95. Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J, Libri D (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell 23: 853–864. doi:10.1016/j.molcel.2006.07.029
96. Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, et al. (2007) Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. Cell 131: 1340–1353. doi:10.1016/j.cell.2007.10.056
97. Vasiljeva L, Kim M, Terzi N, Soares LM, Buratowski S (2008) Transcription termination and RNA degradation contribute to silencing of RNA polymerase II transcription within heterochromatin. Mol Cell 29: 313–323. doi:10.1016/j.molcel.2008.01.011
98. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, et al. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. Science 322: 1851–1854. doi:10.1126/science.1164096
99. Milligan L, Decourty L, Saveanu C, Rappsilber J, Ceulemans H, et al. (2008) A yeast exosome cofactor, Mpp6, functions in RNA surveillance and in the degradation of noncoding RNA transcripts. Mol Cell Biol 28: 5446–5457. doi:10.1128/MCB.00463-08
100. Neil H, Malabat C, d' Aubenton-Carafa Y, Xu Z, Steinmetz LM, et al. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature 457: 1038–1042. doi:10.1038/nature07747
101. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, et al. (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature 457: 1033–1037. doi:10.1038/nature07728
102. Masuda S, Das R, Cheng H, Hurt E, Dorman N, et al. (2005) Recruitment of the human TREX complex to mRNA during splicing. Genes Dev 19: 1512–1517. doi:10.1101/gad.1302205
103. Cheng H, Dufu K, Lee C-S, Hsu JL, Dias A, et al. (2006) Human mRNA export machinery recruited to the 5′ end of mRNA. Cell 127: 1389–1400. doi:10.1016/j.cell.2006.10.044
104. Luo MJ, Reed R (1999) Splicing is required for rapid and efficient mRNA export in metazoans. Proc Natl Acad Sci U S A 96: 14937–14942.
105. Palazzo AF, Springer M, Shibata Y, Lee C-S, Dias AP, et al. (2007) The signal sequence coding region promotes nuclear export of mRNA. PLoS Biol 5: e322. doi:10.1371/journal.pbio.0050322
106. Valencia P, Dias AP, Reed R (2008) Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc Natl Acad Sci U S A 105: 3386–3391. doi:10.1073/pnas.0800250105
107. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. Nature 416: 499–506. doi:10.1038/416499a
108. Buratowski S (2009) Progression through the RNA polymerase II CTD cycle. Mol Cell 36: 541–546. doi:10.1016/j.molcel.2009.10.019
109. Perales R, Bentley D (2009) "Cotranscriptionality": the transcription elongation complex as a

nexus for nuclear transactions. Mol Cell 36: 178–191. doi:10.1016/j.molcel.2009.09.018
110. Moore MJ, Proudfoot NJ (2009) Pre-mRNA processing reaches back to transcription and ahead to translation. Cell 136: 688–700. doi:10.1016/j.cell.2009.02.001
111. Palazzo AF, Akef A (2012) Nuclear export as a key arbiter of "mRNA identity" in eukaryotes. Biochim Biophys Acta 1819: 566–577. doi:10.1016/j.bbagrm.2011.12.012
112. Palazzo A, Mahadevan K, Tarnawsky S (2013) ALREX-elements and introns: two identity elements that promote mRNA nuclear export. WIREs RNA 4: 523–533. doi:10.1002/wrna.1176
113. Ohno M, Segref A, Kuersten S, Mattaj IW (2002) Identity elements used in export of mRNAs. Mol Cell 9: 659–671.
114. Dias AP, Dufu K, Lei H, Reed R (2010) A role for TREX components in the release of spliced mRNA from nuclear speckle domains. Nat Commun 1: 97. doi:10.1038/ncomms1103
115. Lei H, Dias AP, Reed R (2011) Export and stability of naturally intronless mRNAs require specific coding region sequences and the TREX mRNA export complex. Proc Natl Acad Sci U S A 108: 17985–17990. doi:10.1073/pnas.1113076108
116. Huang Y, Steitz JA (2001) Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. Mol Cell 7: 899–905.
117. Culjkovic B, Topisirovic I, Skrabanek L, Ruiz-Gutierrez M, Borden KLB (2006) eIF4E is a central node of an RNA regulon that governs cellular proliferation. J Cell Biol 175: 415–426. doi:10.1083/jcb.200607020
118. Lei H, Zhai B, Yin S, Gygi S, Reed R (2012) Evidence that a consensus element found in naturally intronless mRNAs promotes mRNA export. Nucleic Acids Res. doi:10.1093/nar/gks1314
119. Kimura T, Hashimoto I, Nishizawa M, Ito S, Yamada H (2010) Novel cis-active structures in the coding region mediate CRM1-dependent nuclear export of IFN-α 1 mRNA. Med Mol Morphol 43: 145–157. doi:10.1007/s00795-010-0492-5
120. Mattick JS, Dinger ME (2013) The extent of functionality in the human genome. HUGO J 7: 2. doi:10.1186/1877-6566-7-2
121. Tisseur M, Kwapisz M, Morillon A (2011) Pervasive transcription - Lessons from yeast. Biochimie 93: 1889–1896. doi:10.1016/j.biochi.2011.07.001
122. Moazed D (2009) Small RNAs in transcriptional gene silencing and genome defence. Nature 457: 413–420. doi:10.1038/nature07756
123. Bartolomei MS, Zemel S, Tilghman SM (1991) Parental imprinting of the mouse H19 gene. Nature 351: 153–155. doi:10.1038/351153a0
124. Kobayashi T, Ganley ARD (2005) Recombination regulation by transcription-induced cohesin dissociation in rDNA repeats. Science 309: 1581–1584. doi:10.1126/science.1116102
125. Tan-Wong SM, Zaugg JB, Camblong J, Xu Z, Zhang DW, et al. (2012) Gene loops enhance transcriptional directionality. Science 338: 671–675. doi:10.1126/science.1224350
126. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long Noncoding RNAs with Enhancer-like Function in Human Cells. Cell 143: 46–58. doi:10.1016/j.cell.2010.09.001.
127. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. (2014) An atlas of active enhancers across human cell types and tissues. Nature 507: 455–461. doi:10.1038/nature12787.
128. Bird A (2013) Genome biology: not drowning but waving. Cell 154: 951–952. doi:10.1016/j.cell.2013.08.010
129. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628. doi:10.1038/nmeth.1226

D.7 Expression analysis

Further explanations for Chapter 7.

D.7.1 The first genome-wide methods: microarrays

- Lowe et al. (2017) — A short primer that compares microarrays and RNA-seq side by side, from sample preparation and normalisation to differential expression. Useful for seeing why RNA-seq replaced arrays and what the two technologies share. *(primer; CC BY 4.0)*

  Transcriptomics technologies. Lowe et al. (2017). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1005457

  > Transcriptomics technologies are the techniques used to study an organism's transcriptome, the sum of all of its RNA transcripts. The information content of an organism is recorded in the DNA of its genome and expressed through transcription. Here, mRNA serves as a transient intermediary molecule in the information network, whilst noncoding RNAs perform additional diverse functions. A transcriptome captures a snapshot in time of the total transcripts present in a cell. The first attempts to study the whole transcriptome began in the early 1990s, and technological advances since the late 1990s have made transcriptomics a widespread discipline. Transcriptomics has been defined by repeated technological innovations that transform the field. There are two key contemporary techniques in the field: microarrays, which quantify a set of predetermined sequences, and RNA sequencing (RNA-Seq), which uses high-throughput sequencing to capture all sequences. Measuring the expression of an organism's genes in different tissues, conditions, or time points gives information on how genes are regulated and reveals details of an organism's biology. It can also help to infer the functions of previously unannotated genes. Transcriptomic analysis has enabled the study of how gene expression changes in different organisms and has been instrumental in the understanding of human disease. An analysis of gene expression in its entirety allows detection of broad coordinated trends which cannot be discerned by more targeted assays.

- Slonim and Yanai (2009) — An introductory tutorial on the microarray analysis workflow: quality control, normalisation, differential expression, clustering and functional interpretation. Older, but it describes the classical array pipeline that the section summarises. *(tutorial; CC BY 4.0)*

  Getting Started in Gene Expression Microarray Analysis. Slonim & Yanai (2009). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000543

  > Gene expression microarrays provide a snapshot of all the transcriptional activity in a biological sample. Unlike most traditional molecular biology tools, which generally allow the study of a single gene or a small set of genes, microarrays facilitate the discovery of totally novel and unexpected functional roles of genes. The power of these tools has been applied to a range of applications, including discovering novel disease subtypes, developing new diagnostic tools, and identifying underlying mechanisms of disease or drug response. However, this technology necessarily produces a large amount of data, challenging us to interpret it by exploiting modern computational and statistical tools. In this brief review, we aim to indicate the major issues involved in microarray analysis and provide a useful starting point for new microarray users. Figure 1 outlines the steps in a typical expression microarray experiment and maps them to the different sections of this review. Figure 1 Overview of steps in a typical gene expression microarray experiment.

D.7.2 RNA sequencing

- Law et al. (2018) — A complete worked analysis of a real RNA-seq data set in R, from count filtering and TMM normalisation to MDS plots, voom/limma testing and gene set tests. Shows in practice how the normalisation and modelling steps of the section fit together. *(tutorial; CC BY 4.0)*

  RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. Law et al. (2018). *F1000Research*. doi:10.12688/f1000research.9005.3

> The ability to easily and efficiently analyse RNA-sequencing data is a key strength of the Bioconductor project. Starting with counts summarised at the gene-level, a typical analysis involves pre-processing, exploratory data analysis, differential expression testing and pathway analysis with the results obtained informing future experiments and validation studies. In this workflow article, we analyse RNA-sequencing data from the mouse mammary gland, demonstrating use of the popular edgeR package to import, organise, filter and normalise the data, followed by the limma package with its voom method, linear modelling and empirical Bayes moderation to assess differential expression and perform gene set testing. This pipeline is further enhanced by the Glimma package which enables interactive exploration of the results so that individual samples and genes can be examined by the user. The complete analysis offered by these three packages highlights the ease with which researchers can turn the raw counts from an RNA-sequencing experiment into biological insights using Bioconductor.

- Love et al. (2016) — A Bioconductor workflow from transcript quantification to DESeq2 testing, including size-factor normalisation, variance-stabilising transformations, exploratory PCA and batch handling. A useful complement to the limma/edgeR workflow above. *(tutorial; CC BY 4.0)*

  RNA-Seq workflow: gene-level exploratory analysis and differential expression. Love et al. (2016). *F1000Research*. doi:10.12688/f1000research.7035.2

  > Here we walk through an end-to-end gene-level RNA-Seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were aligned to the reference genome, and prepare a count matrix which tallies the number of RNA-seq reads/fragments within each gene for each sample.We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

D.7.3 Testing and statistics

- Korthauer et al. (2019) — Explains and benchmarks Bonferroni, Benjamini–Hochberg, Storey's q-value and newer covariate-aware FDR methods on genomic examples. Helpful for understanding the multiple-testing part of the section and when the methods differ. *(review; CC BY 4.0)*

  A practical guide to methods controlling false discoveries in computational biology. Korthauer et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1716-1

  > Background In high-throughput studies, hundreds to millions of hypotheses are typically tested. Statistical methods that control the false discovery rate (FDR) have emerged as popular and powerful tools for error rate control. While classic FDR methods use only p values as input, more modern FDR methods have been shown to increase power by incorporating complementary information as informative covariates to prioritize, weight, and group hypotheses. However, there is currently no consensus on how the modern methods compare to one another. We investigate the accuracy, applicability, and ease of use of two classic and six modern FDR-controlling methods by performing a systematic benchmark comparison using simulation studies as well as six case studies in computational biology. Results Methods that incorporate informative covariates are modestly more powerful than classic approaches, and do not underperform classic approaches, even when the covariate is completely uninformative. The majority of methods are successful at controlling the FDR, with the exception of two modern methods under certain settings. Furthermore, we find that the improvement of the modern FDR methods over the classic methods increases with the informativeness of the covariate, total number of hypothesis tests, and proportion of truly non-null hypotheses. Conclusions Modern FDR methods that use an informative covariate provide advantages over classic FDR-controlling procedures, with the relative gain dependent on the application and informativeness of available covariates. We present our findings as a practical guide and provide recommendations to aid researchers in their choice of methods to correct for false discoveries.

- Nguyen and Holmes (2019) — Short practical advice on PCA and related methods: preprocessing, choosing the number of dimensions, interpreting axes and checking stability. Supports the clustering and dimensionality-reduction subsection. *(tutorial; CC BY 4.0)*

  Ten quick tips for effective dimensionality reduction. Nguyen & Holmes (2019). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1006907

- Kobak and Berens (2019) — Explains how t-SNE works and how initialisation, perplexity and exaggeration affect the global structure of the embedding. Useful background for the warnings in the section about over-interpreting t-SNE and UMAP plots. *(tutorial; CC BY 4.0)*

  The art of using t-SNE for single-cell transcriptomics. Kobak & Berens (2019). *Nature Communications*. doi:10.1038/s41467-019-13056-x

  > Single-cell transcriptomics yields ever growing data sets containing RNA expression levels for thousands of genes from up to millions of cells. Common data analysis pipelines include a dimensionality reduction step for visualising the data in two dimensions, most frequently performed using t-distributed stochastic neighbour embedding (t-SNE). It excels at revealing local structure in high-dimensional data, but naive applications often suffer from severe shortcomings, e.g. the global structure of the data is not represented accurately. Here we describe how to circumvent such pitfalls, and develop a protocol for creating more faithful t-SNE visualisations. It includes PCA initialisation, a high learning rate, and multi-scale similarity kernels; for very large data sets, we additionally use exaggeration and downsampling-based initialisation. We use published single-cell RNA-seq data sets to demonstrate that this protocol yields superior results compared to the naive application of t-SNE.

D.7.4 From gene lists to biology

- Wijesooriya et al. (2022) — Surveys how enrichment analysis is used in published studies and lists common errors, such as an inappropriate background gene list and missing correction for multiple testing. A practical guide to doing over-representation analysis correctly. *(perspective; CC BY 4.0)*

  Urgent need for consistent standards in functional enrichment analysis. Wijesooriya et al. (2022). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1009935

  > Gene set enrichment tests (a.k.a. functional enrichment analysis) are among the most frequently used methods in computational biology. Despite this popularity, there are concerns that these methods are being applied incorrectly and the results of some peer-reviewed publications are unreliable. These problems include the use of inappropriate background gene lists, lack of false discovery rate correction and lack of methodological detail. To ascertain the frequency of these issues in the literature, we performed a screen of 186 open-access research articles describing functional enrichment results. We find that 95% of analyses using over-representation tests did not implement an appropriate background gene list or did not describe this in the methods. Failure to perform p-value correction for multiple tests was identified in 43% of analyses. Many studies lacked detail in the methods section about the tools and gene sets used. An extension of this survey showed that these problems are not associated with journal or article level bibliometrics. Using seven independent RNA-seq datasets, we show misuse of enrichment tools alters results substantially. In conclusion, most published functional enrichment studies suffered from one or more major flaws, highlighting the need for stronger standards for enrichment analysis.

- Dam et al. (2017) — Reviews how co-expression networks are built and used to predict gene function and disease genes, including guilt by association, WGCNA-type modules and data requirements. Covers the co-expression part of the section. *(review; CC BY 4.0)*

  Gene co-expression analysis for functional classification and gene–disease predictions. van Dam et al. (2017). *Briefings in Bioinformatics*. doi:10.1093/bib/bbw139

  > Gene co-expression networks can be used to associate genes of unknown function with biological processes, to prioritize candidate disease genes or to discern transcriptional regulatory programmes. With recent advances in transcriptomics and next-generation sequencing, co-expression networks constructed from RNA sequencing data also enable the inference of functions and disease associations for non-coding genes and splice variants. Although gene co-expression networks typically do not provide information about causality, emerging methods for differential co-expression analysis are enabling the identification of regulatory genes underlying various phenotypes. Here, we introduce and guide researchers through a (differential) co-expression analysis. We provide an overview of methods and tools used to create and analyse co-expression networks constructed from gene expression data, and we explain how these can be used to identify genes with a regulatory role in disease. Furthermore, we discuss the integration of other data types with co-expression networks and offer future perspectives of co-expression analysis.

- Gibson et al. (2015) — Introduces cis- and trans-eQTL mapping and how eQTLs are used to interpret disease-associated variants from GWAS. Background for the regulatory-variation paragraph. *(review; CC BY 4.0)*

  Expression quantitative trait locus analysis for translational medicine. Gibson et al. (2015). *Genome Medicine*. doi:10.1186/s13073-015-0186-7

  > Expression quantitative trait locus analysis has emerged as an important component of efforts to understand how genetic polymorphisms influence disease risk and is poised to make contributions to translational medicine. Here we review how expression quantitative trait locus analysis is aiding the identification of which gene(s) within regions of association are causal for a disease or phenotypic trait; the narrowing down of the cell types or regulators involved in the etiology of disease; the characterization of drivers and modifiers of cancer; and our understanding of how different environments and cellular contexts can modify gene expression. We also introduce the concept of transcriptional risk scores as a means of refining estimates of individual liability to disease based on targeted profiling of the transcripts that are regulated by polymorphisms jointly associated with disease and gene expression.

D.7.5 Single-cell genomics

- Haque et al. (2017) — An accessible overview of single-cell RNA-seq: cell isolation, plate- and droplet-based protocols, UMIs, and the main analysis steps. Good for the technologies and properties-of-the-data subsections; the analysis tools have since moved on. *(review; CC BY 4.0)*

  A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Haque et al. (2017). *Genome Medicine*. doi:10.1186/s13073-017-0467-4

  > RNA sequencing (RNA-seq) is a genomic approach for the detection and quantitative analysis of messenger RNA molecules in a biological sample and is useful for studying cellular responses. RNA-seq has fueled much discovery and innovation in medicine over recent years. For practical reasons, the technique is usually conducted on samples comprising thousands to millions of cells. However, this has hindered direct assessment of the fundamental unit of biology-the cell. Since the first single-cell RNA-sequencing (scRNA-seq) study was published in 2009, many more have been conducted, mostly by specialist laboratories with unique skills in wet-lab single-cell genomics, bioinformatics, and computation. However, with the increasing commercial availability of scRNA-seq platforms, and the rapid ongoing maturation of bioinformatics approaches, a point has been reached where any biomedical researcher or clinician can use scRNA-seq to make exciting discoveries. In this review, we present a practical guide to help researchers design their first scRNA-seq studies, including introductory information on experimental hardware, protocol choice, quality control, data analysis and biological interpretation.

- Williams et al. (2022) — Describes sequencing-based and imaging-based spatial transcriptomics methods, their resolution and gene coverage, and the computational analyses applied to them. Covers the spatial transcriptomics subsection. *(review; CC BY 4.0)*

  An introduction to spatial transcriptomics for biomedical research. Williams et al. (2022). *Genome Medicine*. doi:10.1186/s13073-022-01075-1

  > Single-cell transcriptomics (scRNA-seq) has become essential for biomedical research over the past decade, particularly in developmental biology, cancer, immunology, and neuroscience. Most commercially available scRNA-seq protocols require cells to be recovered intact and viable from tissue. This has precluded many cell types from study and largely destroys the spatial context that could otherwise inform analyses of cell identity and function. An increasing number of commercially available platforms now facilitate spatially resolved, high-dimensional assessment of gene transcription, known as 'spatial transcriptomics'. Here, we introduce different classes of method, which either record the locations of hybridized mRNA molecules in tissue, image the positions of cells themselves prior to assessment, or employ spatial arrays of mRNA probes of pre-determined location. We review sizes of tissue area that can be assessed, their spatial resolution, and the number and types of genes that can be profiled. We discuss if tissue preservation influences choice of platform, and provide guidance on whether specific platforms may be better suited to discovery screens or hypothesis testing. Finally, we introduce bioinformatic methods for analysing spatial transcriptomic data, including pre-processing, integration with existing scRNA-seq data, and inference of cell-cell interactions. Spatial -omics methods are already improving our

understanding of human tissues in research, diagnostic, and therapeutic settings. To build upon these recent advancements, we provide entry-level guidance for those seeking to employ spatial transcriptomics in their own biomedical research.

D.7.6 Single-cell sequencing and cancer evolution

- Vendramin et al. (2021) — Reviews cancer as an evolutionary process: clonal selection, intratumour heterogeneity and genetic as well as non-genetic sources of diversity. Background for the clonal-evolution subsection. *(review; CC BY 4.0)*

  Cancer evolution: Darwin and beyond. Vendramin et al. (2021). *The EMBO Journal*. doi:10.15252/embj.2021108389

  > Clinical and laboratory studies over recent decades have established branched evolution as a feature of cancer. However, while grounded in somatic selection, several lines of evidence suggest a Darwinian model alone is insufficient to fully explain cancer evolution. First, the role of macroevolutionary events in tumour initiation and progression contradicts Darwin's central thesis of gradualism. Whole-genome doubling, chromosomal chromoplexy and chromothripsis represent examples of single catastrophic events which can drive tumour evolution. Second, neutral evolution can play a role in some tumours, indicating that selection is not always driving evolution. Third, increasing appreciation of the role of the ageing soma has led to recent generalised theories of age-dependent carcinogenesis. Here, we review these concepts and others, which collectively argue for a model of cancer evolution which extends beyond Darwin. We also highlight clinical opportunities which can be grasped through targeting cancer vulnerabilities arising from non-Darwinian patterns of evolution.

- Beerenwinkel et al. (2014) — Reviews population-genetic models of tumour growth and the computational inference of clonal composition and tumour phylogenies from sequencing data. Links the section to the phylogenetic methods of the book. *(review; CC BY 4.0)*

  Cancer Evolution: Mathematical Models and Computational Inference. Beerenwinkel et al. (2014). *Systematic Biology*. doi:10.1093/sysbio/syu081

  > Cancer is a somatic evolutionary process characterized by the accumulation of mutations, which contribute to tumor growth, clinical progression, immune escape, and drug resistance development. Evolutionary theory can be used to analyze the dynamics of tumor cell populations and to make inference about the evolutionary history of a tumor from molecular data. We review recent approaches to modeling the evolution of cancer, including population dynamics models of tumor initiation and progression, phylogenetic methods to model the evolutionary relationship between tumor subclones, and probabilistic graphical models to describe dependencies among mutations. Evolutionary modeling helps to understand how tumors arise and will also play an increasingly important prognostic role in predicting disease progression and the outcome of medical interventions, such as targeted therapy.

- Navin (2015) — Reviews early single-cell DNA sequencing of tumours, including whole-genome amplification, copy-number profiling and the punctuated evolution observed in breast cancer. Directly relevant to the single-cell DNA sequencing subsection; link only because of the NC licence. *(review; CC BY-NC (link only))*

  The first five years of single-cell cancer genomics and beyond. Navin (2015). *Genome Research*. doi:10.1101/gr.191098.115

  > Single-cell sequencing (SCS) is a powerful new tool for investigating evolution and diversity in cancer and understanding the role of rare cells in tumor progression. These methods have begun to unravel key questions in cancer biology that have been difficult to address with bulk tumor measurements. Over the past five years, there has been extraordinary progress in technological developments and research applications, but these efforts represent only the tip of the iceberg. In the coming years, SCS will greatly improve our understanding of invasion, metastasis, and therapy resistance during cancer progression. These tools will also have direct translational applications in the clinic, in areas such as early detection, noninvasive monitoring, and guiding targeted therapy. In this perspective, I discuss the progress that has been made and the myriad of unexplored applications that still lie ahead in cancer research and medicine.

The following open-access paper(s) are reproduced under their original licences.

PLOS | COMPUTATIONAL BIOLOGY

TOPIC PAGE

Transcriptomics technologies

Rohan Lowe[1], Neil Shirley[2], Mark Bleackley[1], Stephen Dolan[3], Thomas Shafee[1]*

1 La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia, **2** ARC Centre of Excellence in Plant Cell Walls, University of Adelaide, Adelaide, Australia, **3** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

* T.Shafee@LaTrobe.edu.au

Abstract

Transcriptomics technologies are the techniques used to study an organism's transcriptome, the sum of all of its RNA transcripts. The information content of an organism is recorded in the DNA of its genome and expressed through transcription. Here, mRNA serves as a transient intermediary molecule in the information network, whilst noncoding RNAs perform additional diverse functions. A transcriptome captures a snapshot in time of the total transcripts present in a cell.

The first attempts to study the whole transcriptome began in the early 1990s, and technological advances since the late 1990s have made transcriptomics a widespread discipline. Transcriptomics has been defined by repeated technological innovations that transform the field. There are two key contemporary techniques in the field: microarrays, which quantify a set of predetermined sequences, and RNA sequencing (RNA-Seq), which uses high-throughput sequencing to capture all sequences.

Measuring the expression of an organism's genes in different tissues, conditions, or time points gives information on how genes are regulated and reveals details of an organism's biology. It can also help to infer the functions of previously unannotated genes. Transcriptomic analysis has enabled the study of how gene expression changes in different organisms and has been instrumental in the understanding of human disease. An analysis of gene expression in its entirety allows detection of broad coordinated trends which cannot be discerned by more targeted assays.

Check for updates

OPEN ACCESS

Citation: Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T (2017) Transcriptomics technologies. PLoS Comput Biol 13(5): e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Published: May 18, 2017

Copyright: © 2017 Lowe et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Australian Research Council grant DP160100309. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a "Topic Page" article for *PLOS Computational Biology*.

History

Transcriptomics has been characterised by the development of new techniques which have redefined what is possible every decade or so and render previous technologies obsolete (Fig 1). The first attempt at capturing a partial human transcriptome was published in 1991 and

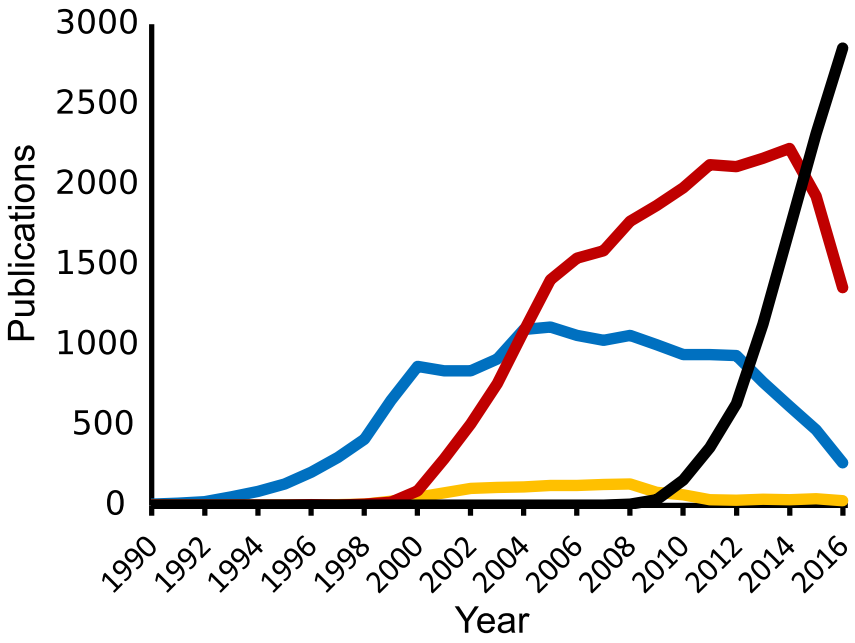


Fig 1. Transcriptomics method use over time. Published papers since 1990, referring to RNA sequencing (black), RNA microarray (red), expressed sequence tag (blue), and serial/cap analysis of gene expression (yellow)[12].

https://doi.org/10.1371/journal.pcbi.1005457.g001

reported 609 mRNA sequences from the human brain [1]. In 2008, two human transcriptomes composed of millions of transcript-derived sequences covering 16,000 genes were published [2][3], and, by 2015, transcriptomes had been published for hundreds of individuals [4][5]. Transcriptomes of different disease states, tissues, or even single cells are now routinely generated [5][6][7]. This explosion in transcriptomics has been driven by the rapid development of new technologies with an improved sensitivity and economy (Table 1) [8][9][10][11].

Table 1. Comparison of contemporary methods [23] [24] [10].

Method	RNA-Seq	Microarray
Throughput	High [10]	Higher [10]
Input RNA amount	Low ~ 1 ng total RNA [25]	High ~ 1 μg mRNA [26]
Labour intensity	High (sample preparation and data analysis) [10][23]	Low [10][23]
Prior knowledge	None required, though genome sequence useful [23]	Reference transcripts required for probes [23]
Quantitation accuracy	~90% (limited by sequence coverage) [27]	>90% (limited by fluorescence detection accuracy) [27]
Sequence resolution	Can detect SNPs and splice variants (limited by sequencing accuracy of ~99%) [27]	Dedicated arrays can detect splice variants (limited by probe design and cross-hybridisation) [27]
Sensitivity	10^{-6} (limited by sequence coverage) [27]	10^{-3} (limited by fluorescence detection) [27]
Dynamic range	$>10^5$ (limited by sequence coverage) [28]	10^3–10^4 (limited by fluorescence saturation) [28]
Technical reproducibility	>99% [29][30]	>99% [31][32]

RNA-Seq, RNA Sequencing

https://doi.org/10.1371/journal.pcbi.1005457.t001

Before transcriptomics

Studies of individual transcripts were being performed several decades before any transcriptomics approaches were available. Libraries of silkmoth mRNAs were collected and converted to complementary DNA (cDNA) for storage using reverse transcriptase in the late 1970s [13]. In the 1980s, low-throughput Sanger sequencing began to be used to sequence random individual transcripts from these libraries, called expressed sequence tags (ESTs) [2][14][15][16]. The Sanger method of sequencing was predominant until the advent of high-throughput methods such as sequencing by synthesis (Solexa/Illumina, San Diego, CA). ESTs came to prominence during the 1990s as an efficient method to determine the gene content of an organism without sequencing the entire genome [16]. Quantification of individual transcripts by northern blotting, nylon membrane arrays, and later reverse transcriptase quantitative PCR (RT-qPCR) were also popular [17][18], but these methods are laborious and can only capture a tiny subsection of a transcriptome [12]. Consequently, the manner in which a transcriptome as a whole is expressed and regulated remained unknown until higher-throughput techniques were developed.

Early attempts

The word "transcriptome" was first used in the 1990s [19][20]. In 1995, one of the earliest sequencing-based transcriptomic methods was developed, serial analysis of gene expression (SAGE), which worked by Sanger sequencing of concatenated random transcript fragments [21]. Transcripts were quantified by matching the fragments to known genes. A variant of SAGE using high-throughput sequencing techniques, called digital gene expression analysis, was also briefly used [9][22]. However, these methods were largely overtaken by high throughput sequencing of entire transcripts, which provided additional information on transcript structure, e.g., splice variants [9].

Development of contemporary techniques

The dominant contemporary techniques, microarrays and RNA-Seq, were developed in the mid-1990s and 2000s [9][33]. Microarrays that measure the abundances of a defined set of transcripts via their hybridisation to an array of complementary probes were first published in 1995 [34][35]. Microarray technology allowed the assay of thousands of transcripts simultaneously at a greatly reduced cost per gene and labour saving [36]. Both spotted oligonucleotide arrays and Affymetrix (Santa Clara, California) high-density arrays were the method of choice for transcriptional profiling until the late 2000s [12][33]. Over this period, a range of microarrays were produced to cover known genes in model or economically important organisms. Advances in design and manufacture of arrays improved the specificity of probes and allowed for more genes to be tested on a single array. Advances in fluorescence detection increased the sensitivity and measurement accuracy for low abundance transcripts [35][37].

RNA-Seq refers to the sequencing of transcript cDNAs, in which abundance is derived from the number of counts from each transcript. The technique has therefore been heavily influenced by the development of high-throughput sequencing technologies [9][11]. Massively parallel signature sequencing (MPSS) was an early example based on generating 16–20 bp sequences via a complex series of hybridisations [38] and was used in 2004 to validate the expression of 10^4 genes in *Arabidopsis thaliana* [39]. The earliest RNA-Seq work was published in 2006 with 10^5 transcripts sequenced using the 454 technology [40]. This was sufficient coverage to quantify relative transcript abundance. RNA-Seq began to increase in popularity after 2008 when new Solexa/Illumina technologies (San Diego, CA) allowed 10^9 transcript

sequences to be recorded [4][10][41][42]. This yield is now sufficient for accurate quantitation of entire human transcriptomes.

Data gathering

Generating data on RNA transcripts can be achieved via either of two main principles: sequencing of individual transcripts (ESTs, or RNA-Seq), or hybridisation of transcripts to an ordered array of nucleotide probes (i.e., microarrays).

Isolation of RNA

All transcriptomic methods require RNA to first be isolated from the experimental organism before transcripts can be recorded. Although biological systems are incredibly diverse, RNA extraction techniques are broadly similar and involve the following: mechanical disruption of cells or tissues, disruption of RNase with chaotropic salts [43], disruption of macromolecules and nucleotide complexes, separation of RNA from undesired biomolecules including DNA, and concentration of the RNA via precipitation from solution or elution from a solid matrix [43][44]. Isolated RNA may additionally be treated with DNase to digest any traces of DNA [45]. It is necessary to enrich messenger RNA as total RNA extracts are typically 98% ribosomal RNA [46]. Enrichment for transcripts can be performed by poly-A affinity methods or by depletion of ribosomal RNA using sequence-specific probes [47]. Degraded RNA may affect downstream results; for example, mRNA enrichment from degraded samples will result in the depletion of 5′ mRNA ends and uneven signal across the length of a transcript. Snap-freezing of tissue prior to RNA isolation is typical, and care is taken to reduce exposure to RNase enzymes once isolation is complete [44].

EST

An EST is a short nucleotide sequence generated from a single RNA transcript. RNA is first copied as cDNA by a reverse transcriptase enzyme before the resultant cDNA is sequenced [16]. The Sanger method of sequencing was predominant until the advent of high-throughput methods such as sequencing by synthesis (Solexa/Illumina, San Diego, CA). Because ESTs don't require prior knowledge of the organism from which they come, they can also be made from mixtures of organisms or environmental samples [16]. Although higher-throughput methods are now used, EST libraries commonly provided sequence information for early microarray designs; for example, a barley GeneChip was designed from 350,000 previously sequenced ESTs [48].

Serial and Cap analysis of gene expression (SAGE/CAGE)

SAGE was a development of EST methodology to increase the throughput of the tags generated and allow some quantitation of transcript abundance (Fig 2) [21]. cDNA is generated from the RNA but is then digested into 11 bp "tag" fragments using restriction enzymes that cut at a specific sequence, and 11 base pairs along from that sequence. These cDNA tags are then concatenated head-to-tail into long strands (>500 bp) and sequenced using low-throughput, but long read length methods such as Sanger sequencing. Once the sequences are deconvoluted into their original 11 bp tags [21]. If a reference genome is available, these tags can sometimes be aligned to identify their corresponding gene. If a reference genome is unavailable, the tags can simply be directly used as diagnostic markers if found to be differentially expressed in a disease state.

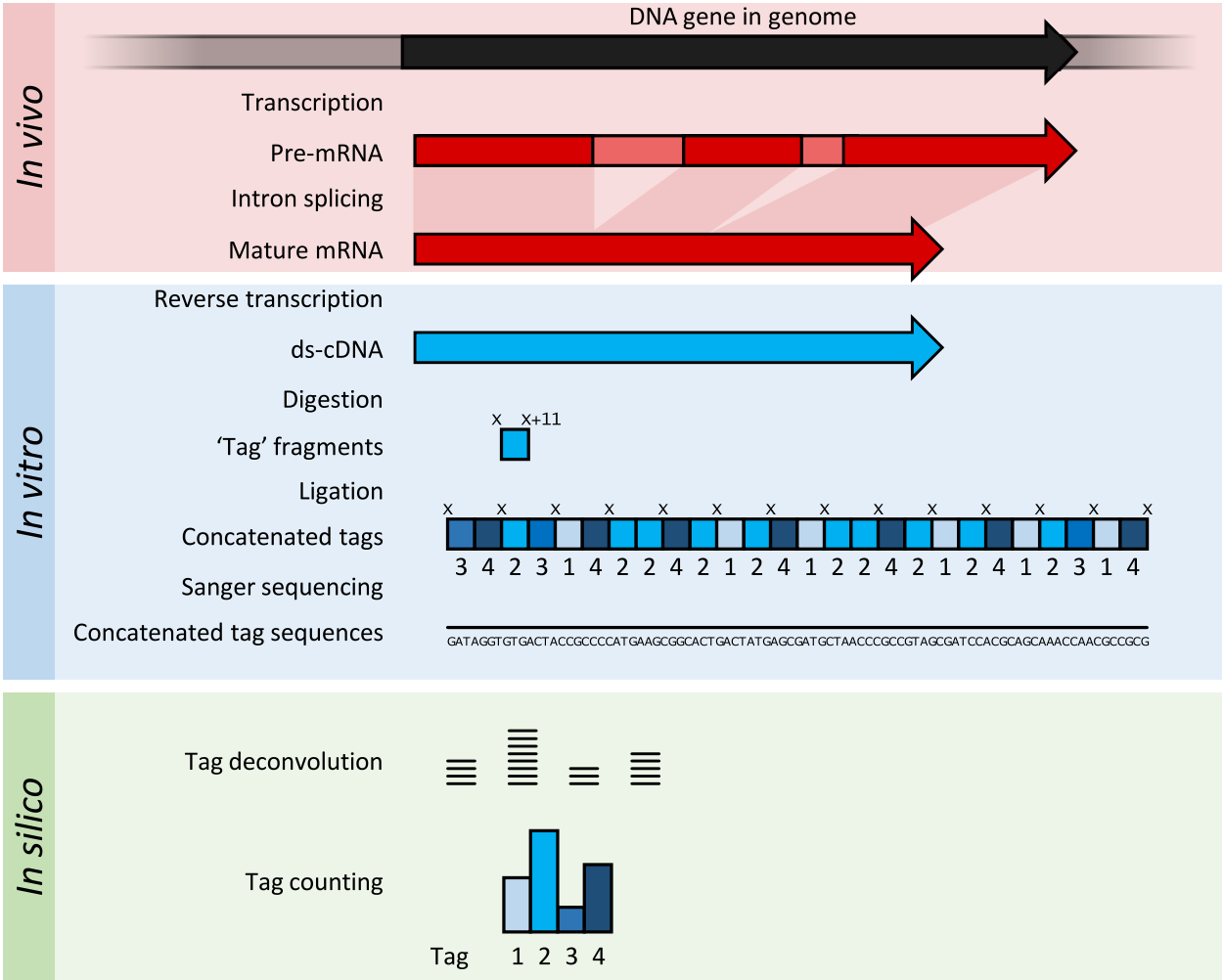


Fig 2. Summary of SAGE. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, and reverse transcriptase is used to copy the mRNA into stable double-stranded–cDNA (ds-cDNA; blue). In SAGE, the ds-cDNA is digested by restriction enzymes (at location "X" and "X"+11) to produce 11-nucleotide "tag" fragments. These tags are concatenated and sequenced using long-read Sanger sequencing (different shades of blue indicate tags from different genes). The sequences are deconvoluted to find the frequency of each tag. The tag frequency can be used to report on transcription of the gene that the tag came from.

https://doi.org/10.1371/journal.pcbi.1005457.g002

The Cap analysis of gene expression (CAGE) method is a variant of SAGE that sequences tags from the 5′ end of an mRNA transcript only [49]. Therefore, the transcriptional start site of genes can be identified when the tags are aligned to a reference genome. Identifying gene start sites is of use for promoter analysis and for the cloning of full-length cDNAs.

SAGE and CAGE methods produce information on more genes than was possible when sequencing single ESTs, but the sample preparation and data analysis are typically more labour intensive.

Microarrays

Principles and advances. Microarrays consist of short nucleotide oligomers, known as "probes," which are arrayed on a solid substrate (e.g., glass) [50]. Transcript abundance is determined by hybridisation of fluorescently labelled transcripts to these probes (Fig 3) [51].

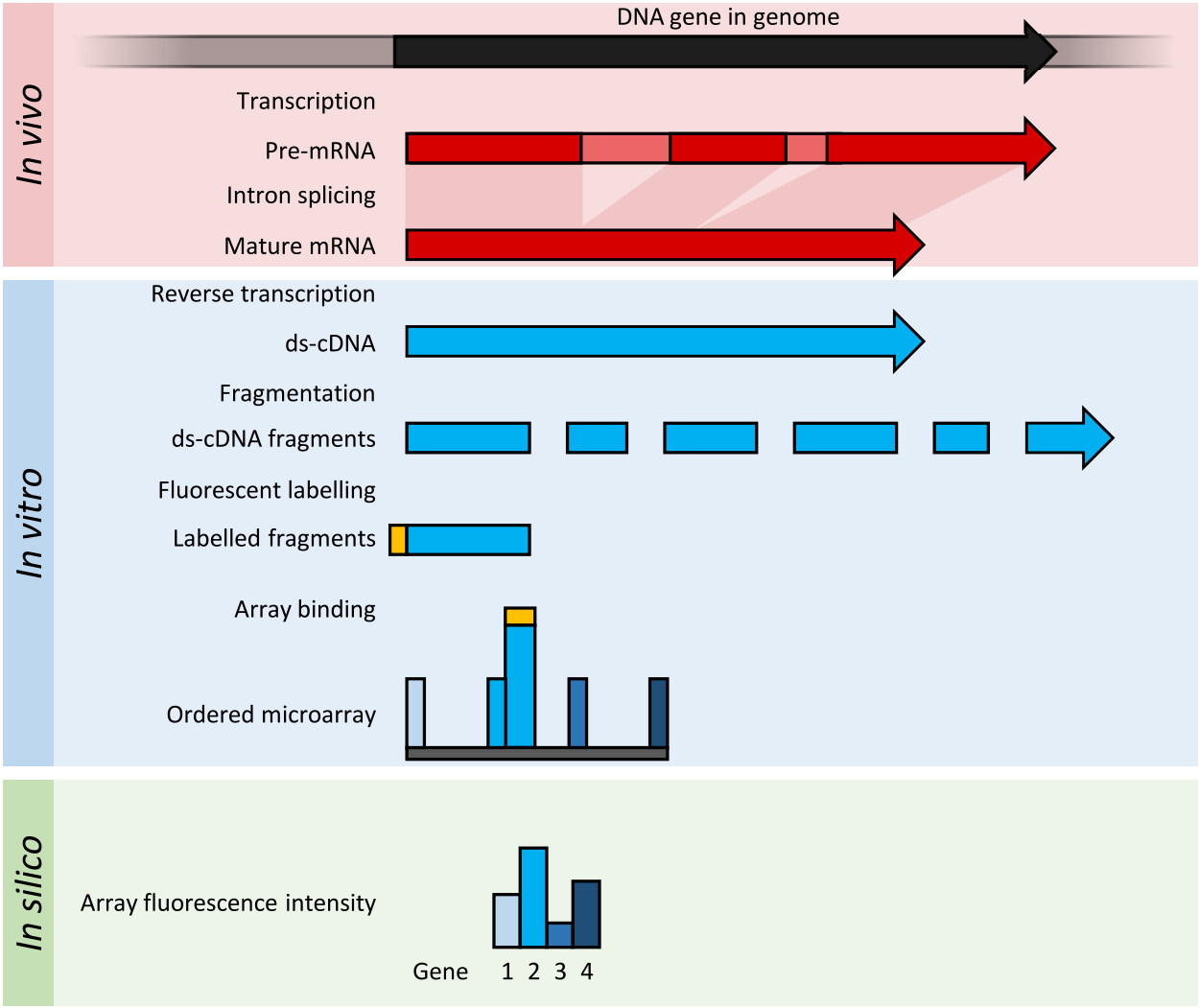


Fig 3. Summary of DNA microarrays. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism and reverse transcriptase is used to copy the mRNA into stable double-stranded–cDNA (ds-cDNA; blue). In microarrays, the ds-cDNA is fragmented and fluorescently labelled (orange). The labelled fragments bind to an ordered array of complementary oligonucleotides, and measurement of fluorescent intensity across the array indicates the abundance of a predetermined set of sequences. These sequences are typically specifically chosen to report on genes of interest within the organism's genome.

https://doi.org/10.1371/journal.pcbi.1005457.g003

The fluorescence intensity at each probe location on the array indicates the transcript abundance for that probe sequence [51]. Microarrays require some prior knowledge of the organism of interest, for example, in the form of an annotated genome sequence or in a library of ESTs that can be used to generate the probes for the array.

Methods. The manufacture of microarrays relies on micro and nanofabrication techniques. Microarrays for transcriptomics typically fall into one of the following two broad categories: low-density spotted arrays or high-density short probe arrays [36]. Transcript presence may be recorded with single- or dual-channel detection of fluorescent tags.

Spotted low-density arrays typically feature picolitre drops of a range of purified cDNAs arrayed on the surface of a glass slide [52]. The probes are longer than those of high-density

arrays and typically lack the transcript resolution of high-density arrays. Spotted arrays use different fluorophores for test and control samples, and the ratio of fluorescence is used to calculate a relative measure of abundance [53]. High-density arrays use single channel detection, and each sample is hybridised and detected individually [54]. High-density arrays were popularised by the Affymetrix GeneChip array (Santa Clara, CA), in which each transcript is quantified by several short 25-mer probes that together assay one gene [55].

NimbleGen arrays (Pleasanton, CA) are high-density arrays produced by a maskless-photochemistry method, which permits flexible manufacture of arrays in small or large numbers. These arrays have hundreds of thousands of 45- to 85-mer probes and are hybridised with a one-colour labelled sample for expression analysis [56]. Some designs incorporate up to 12 independent arrays per slide.

RNA-Seq

Principles and advances. RNA-Seq refers to the combination of a high-throughput sequencing methodology with computational methods to capture and quantify transcripts present in an RNA extract (Fig 4) [10]. The nucleotide sequences generated are typically around 100 bp in length, but can range from 30 bp to over 10,000 bp, depending on the sequencing method used. RNA-Seq leverages deep sampling of the transcriptome with many short fragments from a transcriptome to allow computational reconstruction of the original RNA transcript by aligning reads to a reference genome or to each other (de novo assembly) [9]. The typical dynamic range of 5 orders of magnitude for RNA-Seq is a key advantage over microarray transcriptomes. In addition, input RNA amounts are much lower for RNA-Seq (nanogram quantity) compared to microarrays (microgram quantity), which allowed finer examination of cellular structures, down to the single-cell level when combined with linear amplification of cDNA [25]. Theoretically, there is no upper limit of quantification in RNA-Seq, and background signal is very low for 100 bp reads in nonrepetitive regions [10].

RNA-Seq may be used to identify genes within a genome or identify which genes are active at a particular point in time, and read counts can be used to accurately model the relative gene expression level. RNA-Seq methodology has constantly improved, primarily through the development of DNA sequencing technologies to increase throughput, accuracy, and read length [57]. Since the first descriptions in 2006 and 2008 [40][58], RNA-Seq has been rapidly adopted and overtook microarrays as the dominant transcriptomics technique in 2015 [59].

The quest for transcriptome data at the level of individual cells has driven advances in RNA-Seq library preparation methods, resulting in dramatic advances in sensitivity. Single-cell transcriptomes are now well described and have even been extended to in situ RNA-Seq where transcriptomes of individual cells are directly interrogated in fixed tissues [60].

Methods. RNA-Seq was established in concert with the rapid development of a range of high-throughput DNA sequencing technologies [61]. However, before the extracted RNA transcripts are sequenced, several key processing steps are performed. Methods differ in the use of transcript enrichment, fragmentation, amplification, single or paired-end sequencing, and whether to preserve strand information.

The sensitivity of an RNA-Seq experiment can be increased by enriching classes of RNA that are of interest and depleting known abundant RNAs. The mRNA molecules can be separated by using oligonucleotides probes which bind their poly-A tails. Alternatively, ribo-depletion can be used to specifically remove abundant but uninformative ribosomal RNAs (rRNAs) by hybridisation to probes tailored to the taxon's specific rRNA sequences (e.g., mammal rRNA, plant rRNA). However, ribo-depletion can also introduce some bias via nonspecific

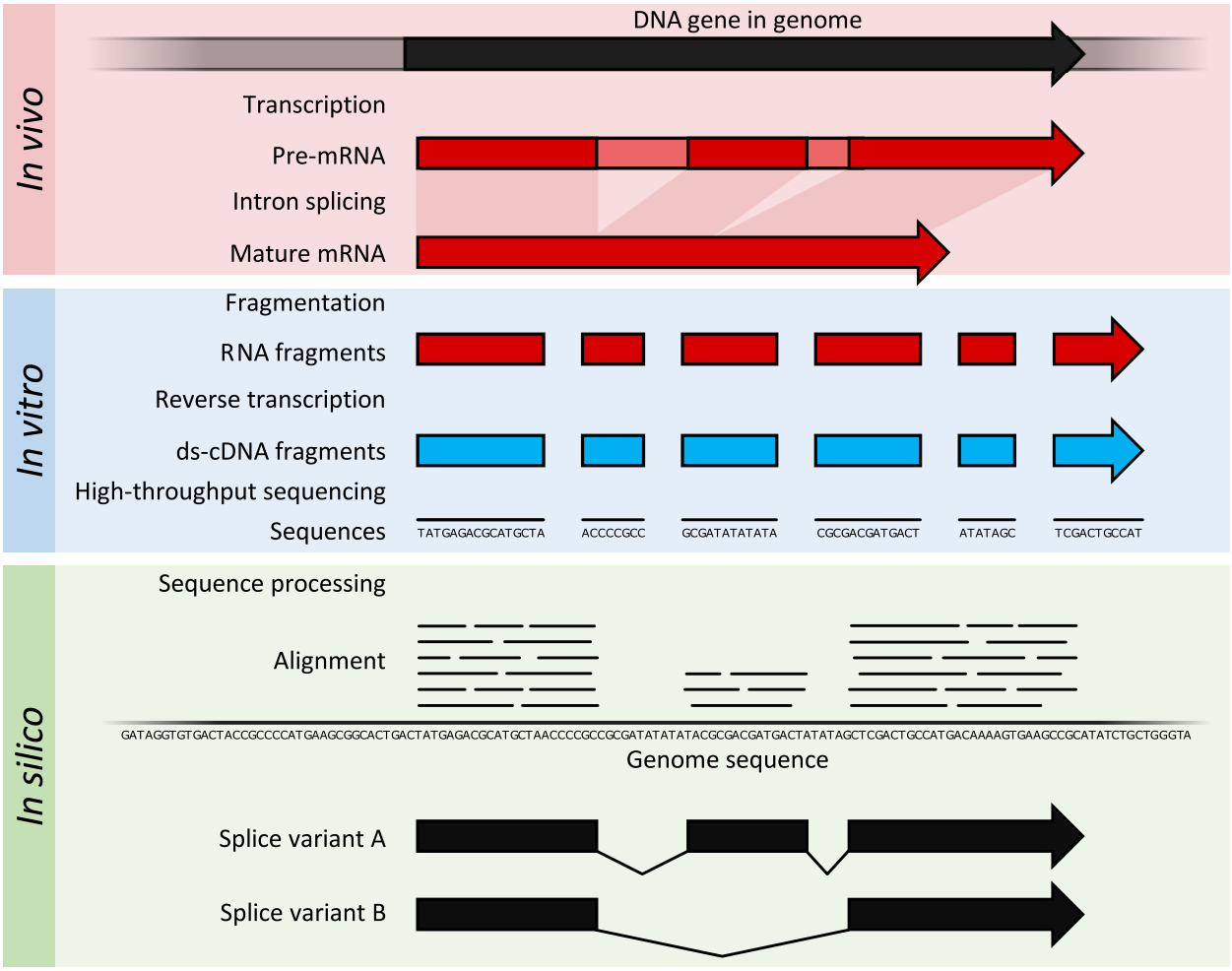


Fig 4. Summary of RNA sequencing. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable double-stranded–cDNA (ds-cDNA; blue). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. These data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.

https://doi.org/10.1371/journal.pcbi.1005457.g004

depletion of off-target transcripts [62]. Small RNAs such as microRNAs, can be purified based on their size by gel electrophoresis and extraction.

Because mRNAs are longer than the read-lengths of typical high-throughput sequencing methods, transcripts are usually fragmented prior to sequencing. The fragmentation method is a key aspect of sequencing library construction [63]. It may incorporate chemical hydrolysis, nebulisation, or sonication of RNA, or utilise simultaneous fragmentation, and tagging of cDNA by transposase enzymes.

During preparation for sequencing, cDNA copies of transcripts may be amplified by PCR to enrich for fragments that contain the expected 5′ and 3′ adapter sequences [64]. Amplification is also used to allow sequencing of very low-input amounts of RNA, down to as little as 50 pg in extreme applications [65]. Spike-in controls can be used to provide quality control assessment of library preparation and sequencing, in terms of guanine-cytosine content, fragment length, as well as the bias due to fragment position within a transcript [66]. Unique molecular

identifiers (UMIs) are short random sequences that are used to individually tag sequence fragments during library preparation so that every tagged fragment is unique [67]. UMIs provide an absolute scale for quantification and the opportunity to correct for subsequent amplification bias introduced during library construction and accurately estimate the initial sample size. UMIs are particularly well-suited to single-cell RNA-Seq transcriptomics, in which the amount of input RNA is restricted and extended amplification of the sample is required [68][69][70].

Once the transcript molecules have been prepared, they can be sequenced in just one direction (single-end) or both directions (paired-end). A single-end sequence is usually quicker to produce, cheaper than paired-end sequencing, and sufficient for quantification of gene expression levels. Paired-end sequencing produces more robust alignments and/or assemblies, which is beneficial for gene annotation and transcript isoform discovery [10]. Strand-specific RNA-Seq methods preserve the strand information of a sequenced transcript [71]. Without strand information, reads can be aligned to a gene locus, but do not inform in which direction the gene is transcribed. Stranded-RNA-Seq is useful for deciphering transcription for genes that overlap in different directions, and to make more robust gene predictions in nonmodel organisms [71].

Currently, RNA-Seq relies on copying of RNA molecules into cDNA molecules prior to sequencing; hence, the subsequent platforms are the same for transcriptomic and genomic data (Table 2). Consequently, the development of DNA sequencing technologies has been a defining feature of RNA-Seq [73][75][76]. Direct sequencing of RNA using nanopore sequencing represents a current state-of-the-art RNA-Seq technique in its infancy (in prerelease beta testing as of 2016) [77][78]. However, nanopore sequencing of RNA can detect modified bases that would be otherwise masked when sequencing cDNA and also eliminates amplification steps that can otherwise introduce bias [11][79].

The sensitivity and accuracy of an RNA-Seq experiment are dependent on the number of reads obtained from each sample. A large number of reads are needed to ensure sufficient coverage of the transcriptome, enabling detection of low abundance transcripts. Experimental design is further complicated by sequencing technologies with a limited output range, the variable efficiency of sequence creation, and variable sequence quality. Added to those considerations is that every species has a different number of genes and therefore requires a tailored sequence yield for an effective transcriptome. Early studies determined suitable thresholds empirically, but as the technology matured, suitable coverage is predicted computationally by transcriptome saturation. Somewhat counterintuitively, the most effective way to improve detection of differential expression in low expression genes is to add more biological replicates, rather than adding

Table 2. Sequencing technology platforms commonly used for RNA-Seq [72][73].

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

https://doi.org/10.1371/journal.pcbi.1005457.t002

more reads [80]. The current benchmarks recommended by the Encyclopedia of DNA Elements (ENCODE) Project are for 70-fold exome coverage for standard RNA-Seq and up to 500-fold exome coverage to detect rare transcripts and isoforms [81][82][83].

Data analysis

Transcriptomics methods are highly parallel and require significant computation to produce meaningful data for both microarray and RNA-Seq experiments. Microarray data are recorded as high-resolution images, requiring feature detection and spectral analysis. Microarray raw image files are each about 750 MB in size, while the processed intensities are around 60 MB in size. Multiple short probes matching a single transcript can reveal details about the intron-exon structure, requiring statistical models to determine the authenticity of the resulting signal. RNA-Seq studies can produce $>10^9$ of short DNA sequences, which must be aligned to reference genomes comprised of millions to billions of base pairs. De novo assembly of reads within a dataset requires the construction of highly complex sequence graphs. RNA-Seq operations are highly repetitious and benefit from parallelised computation, but modern algorithms mean consumer computing hardware is sufficient for simple transcriptomics experiments that do not require de novo assembly of reads. A human transcriptome could be accurately captured by using RNA-Seq with 30 million 100 bp sequences per sample [84][85]. This example would require approximately 1.8 gigabytes of disk space per sample when stored in a compressed fastq format. Processed count data for each gene would be much smaller, equivalent to processed microarray intensities. Sequence data may be stored in public repositories, such as the Sequence Read Archive (SRA) [86]. RNA-Seq datasets can be uploaded via the Gene Expression Omnibus.

Image processing

Microarray image processing must correctly identify the regular grid of features within an image and independently quantify the fluorescence intensity for each feature (Fig 5). Image

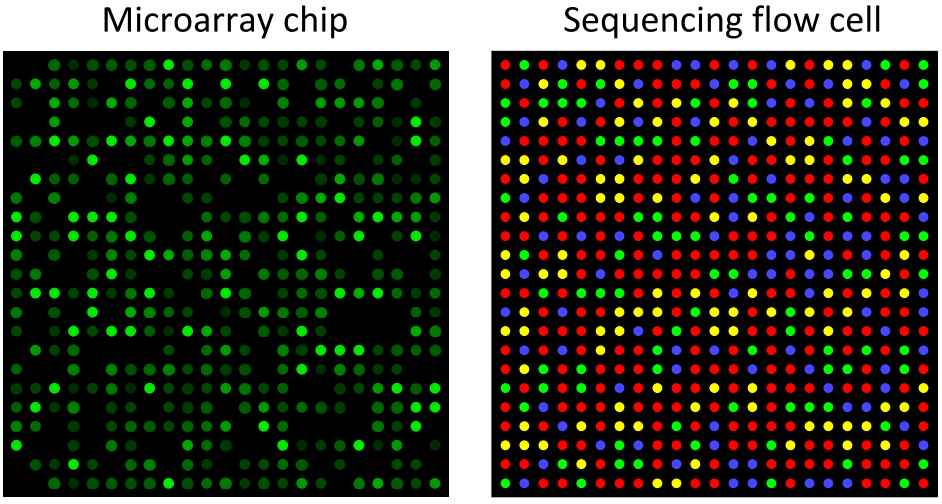


Fig 5. Microarray and sequencing flow cell. Microarrays and RNA sequencing (RNA-Seq) rely on image analysis in different ways. In a microarray chip, each spot on a chip is a defined oligonucleotide probe, and fluorescence intensity directly detects the abundance of a specific sequence (Affymetrix, Santa Clara, CA). In a high-throughput sequencing flow cell, spots are sequenced one nucleotide at a time, with the colour at each round indicating the next nucleotide in the sequence (Illumina Hiseq, San Diego, CA). Other variations of these techniques use more or fewer colour channels.

https://doi.org/10.1371/journal.pcbi.1005457.g005

artefacts must be additionally identified and removed from the overall analysis [87]. Fluorescence intensities directly indicate the abundance of each sequence because the sequence of each probe on the array is already known.

The first steps of RNA-seq also include similar image processing, however conversion of images to sequence data is typically handled automatically by the instrument software. The Illumina sequencing-by-synthesis method results in a random or ordered array of clusters distributed over the surface of a flow cell. The flow cell is imaged up to four times during each sequencing cycle, with tens to hundreds of cycles in total. Flow cell clusters are analogous to microarray spots and must be correctly identified during the early stages of the sequencing process. In Roche's Pyrosequencing method, the intensity of emitted light determines the number of consecutive nucleotides in a homopolymer repeat. There are many variants on these methods, each with a different error profile for the resulting data [88].

RNA-Seq data analysis

RNA-Seq experiments generate a large volume of raw sequence reads, which have to be processed to yield useful information. Data analysis usually requires a combination of bioinformatics software tools that vary according to the experimental design and goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a command-line interface, either in a Unix environment or within the R/Bioconductor statistical environment [90].

Quality control. Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs (k-mers), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the FastQC and FaQCs software packages [91][92]. Abnormalities identified may be removed by trimming or tagged for special treatment during later processes.

Alignment. In order to link sequence read abundance to expression of a particular gene, transcript sequences are aligned to a reference genome, or de novo aligned to one another if no reference is available. The key challenges for alignment software include sufficient speed to permit $>10^9$ of short sequences to be aligned in a meaningful timeframe, flexibility to recognise and deal with intron splicing of eukaryotic mRNA, and correct assignment of reads that map to multiple locations. Software advances have greatly addressed these issues, and increases in sequencing read length are further reducing multimapping reads. A list of currently available high-throughput sequence aligners is maintained by the EBI [93][94].

Alignment of primary transcript mRNA sequences derived from eukaryotes to a reference genome requires specialised handling of intron sequences, which are absent from mature mRNA. Short read aligners perform an additional round of alignments specifically designed to identify splice junctions, informed by canonical splice site sequences and known intron splice site information. Identification of intron splice junctions prevents reads from being misaligned across splice junctions or erroneously discarded, allowing for more reads to be aligned to the reference genome and improving the accuracy of gene expression estimates. Because gene regulation may occur at the mRNA isoform level, splice-aware alignments also permit detection of isoform abundance changes that would otherwise be lost in a bulked analysis [95].

De novo assembly can be used to align reads to one another to construct full-length transcript sequences without the use of a reference genome (Table 3) [96]. Challenges particular to de novo assembly include larger computational requirements compared to a reference-based

Table 3. RNA-Seq de novo assembly software.

Software (Manufacturer)	Released	Last Updated	Resource load	Strengths and weaknesses
Velvet-Oases [100][101]	2008	2011	Heavy	The original short read assembler, now largely superseded.
SOAPdenovo-trans [102]	2011	2015	Moderate	Early short read assembler, updated for transcript assembly.
Trans-ABySS [103]	2010	2016	Moderate	Short reads, large genomes, MPI-parallel version available.
Trinity [104][105]	2011	2017	Moderate	Short reads, large genomes, memory intensive.
miraEST [106]	1999	2016	Moderate	Repetitive sequences, hybrid data input, wide range of sequence platforms accepted.
Newbler [107]	2004	2012	Heavy	Specialised for Roche 454 sequence, homo-polymer error handling.
CLC genomics workbench (Qiagen—Venlo, Netherlands) [108]	2008	2014	Light	Graphical user interface, hybrid data.

MPI, Message Passing Interface; RNA-Seq, RNA sequencing.

https://doi.org/10.1371/journal.pcbi.1005457.t003

transcriptome, additional validation of gene variants or fragments, additional annotation of assembled transcripts. The first metrics used to describe transcriptome assemblies, such as N50, have been shown to be misleading [97], and subsequently improved evaluation methods are now available [98][99]. Annotation-based metrics are better assessments of assembly completeness, such as contig reciprocal best hit count. Once assembled de novo, the assembly can be used as a reference for subsequent sequence alignment methods and quantitative gene expression analysis.

Quantification. Quantification of sequence alignments may be performed at the gene, exon, or transcript level. Typical outputs include a table of reads counts for each feature supplied to the software, for example, for genes in a general feature format file. Gene and exon read counts may be calculated easily using the HTSeq software package, for example [109]. Quantitation at the transcript level is more complicated and requires probabilistic methods to estimate transcript isoform abundance from short read information, for example, using cufflinks software [95]. Reads that align equally well to multiple locations must be identified and either removed, aligned to one of the possible locations, or aligned to the most probable location.

Some quantification methods can circumvent the need for an exact alignment of a read to a reference sequence all together. The kallisto method combines pseudoalignment and quantification into a single step that runs 2 orders of magnitude faster than comparable methods such as tophat/cufflinks, with less computational burden [110].

Differential expression. Once quantitative counts of each transcript are available, differential gene expression is then measured by normalising, modelling, and statistically analysing the data (Fig 6). Examples of dedicated software are described in Table 4. Most read a table of genes and read counts as their input, but some, such as cuffdiff, will accept binary alignment map format read alignments as input. The final outputs of these analyses are gene lists with associated pair-wise tests for differential expression between treatments and the probability estimates of those differences.

Validation

Transcriptomic analyses may be validated using an independent technique, for example, quantitative PCR (qPCR), which is recognisable and statistically assessable [115]. Gene expression is measured against defined standards both for the gene of interest and control genes. The

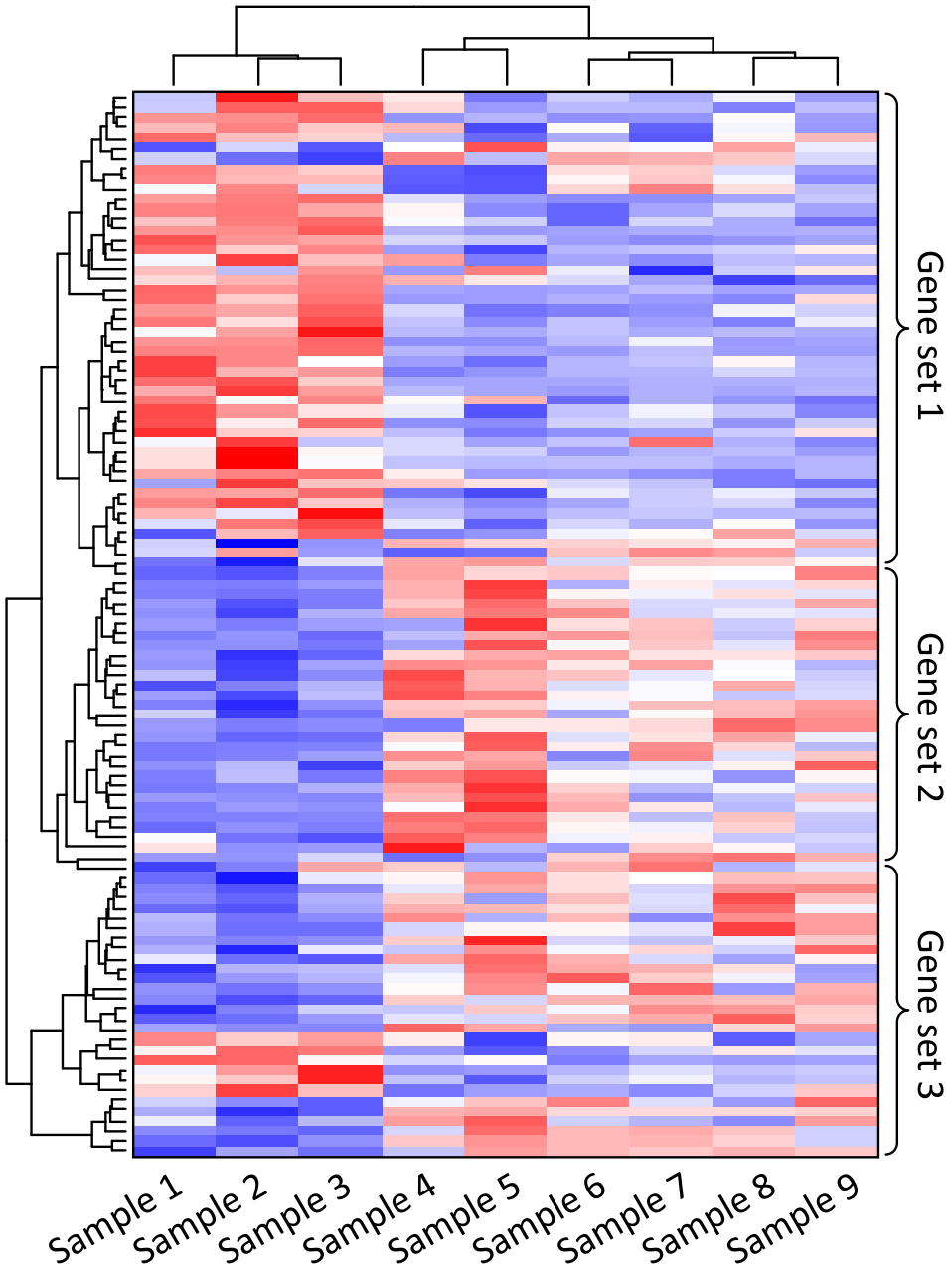


Fig 6. identification of gene co-expression patterns across different samples. Heatmap Each column contains the measurements for gene expression change for a single sample. Relative gene expression is indicated by colour: high-expression (red), median-expression (white) and low-expression (blue). Genes and samples with similar expression profiles can be automatically grouped (left and top trees). Samples may be different individuals, tissues, environments, or health conditions. In this example, expression of gene set 1 is high and expression of gene set 2 is low in samples 1, 2, and 3.

https://doi.org/10.1371/journal.pcbi.1005457.g006

Table 4. RNA-Seq differential gene expression software.

Software	Environment	Specialisation
Cuffdiff2 [111]	Unix-based	Transcript analysis at isoform-level
EdgeR [112]	R/Bioconductor	Any count-based genomic data
DEseq2 [113]	R/Bioconductor	Flexible data types, low replication
Limma/Voom [114]	R/Bioconductor	Microarray or RNA-Seq data, isoform analysis, flexible experiment design

RNA-Seq, RNA sequencing.

https://doi.org/10.1371/journal.pcbi.1005457.t004

measurement by qPCR is similar to that obtained by RNA-Seq wherein a value can be calculated for the concentration of a target region in a given sample. qPCR is, however, restricted to amplicons smaller than 300 bp, usually toward the 3′ end of the coding region, avoiding the 3′ unstralated region (3′UTR) [116]. If validation of transcript isoforms is required, an inspection of RNA-Seq read alignments should indicate where qPCR primers might be placed for maximum discrimination. The measurement of multiple control genes along with the genes of interest produces a stable reference within a biological context [117]. qPCR validation of RNA-Seq data has generally shown that different RNA-Seq methods are highly correlated [58] [118][119].

Functional validation of key genes is an important consideration for post transcriptome planning. Observed gene expression patterns may be functionally linked to a phenotype by an independent knock-down/rescue study in the organism of interest.

Applications

Diagnostics and disease profiling

Transcriptomic strategies have seen broad application across diverse areas of biomedical research, including disease diagnosis and profiling [10]. RNA-Seq approaches have allowed for the large-scale identification of transcriptional start sites and uncovered alternative promoter usage and novel splicing alterations. These regulatory elements are important in human disease, and therefore, defining such variants is crucial to the interpretation of disease-association studies [120]. RNA-Seq can also identify disease-associated single nucleotide polymorphisms (SNP), allele-specific expression, and gene fusions, contributing to our understanding of disease causal variants [121].

Retrotransposons are transposable elements which proliferate within eukaryotic genomes through a process involving reverse transcription. RNA-Seq can provide information about the transcription of endogenous retrotransposons that may influence the transcription of neighbouring genes by various epigenetic mechanisms that lead to disease [122]. Similarly, the potential for using RNA-Seq to understand immune-related disease is expanding rapidly due to the ability to dissect immune cell populations and to sequence T cell and B cell receptor repertoires from patients [123][124].

Human and pathogen transcriptomes

RNA-Seq of human pathogens has become an established method for quantifying gene expression changes, identifying novel virulence factors, predicting antibiotic resistance, and unveiling host-pathogen immune interactions [125][126]. A primary aim of this technology is to develop optimised infection control measures and targeted, individualised treatment [124].

Transcriptomic analysis has predominantly focused on either the host or the pathogen. Dual RNA-Seq has recently been applied to simultaneously profile RNA expression in both the

pathogen and host throughout the infection process. This technique enables the study of the dynamic response and interspecies gene regulatory networks in both interaction partners from initial contact through to invasion and the final persistence of the pathogen or clearance by the host immune system [127][128].

Responses to environment

Transcriptomics allows for the identification of genes and pathways that respond to and counteract biotic and abiotic environmental stresses. The nontargeted nature of transcriptomics allows for the identification of novel transcriptional networks in complex systems. For example, comparative analysis of a range of chickpea lines at different developmental stages identified distinct transcriptional profiles associated with drought and salinity stresses, including identifying the role of transcript isoforms of Apetela 2 and Ethylene-Responsive Element Binding Protein (AP2-EREBP) [129]. Investigation of gene expression during biofilm formation by the fungal pathogen *Candida albicans* revealed a coregulated set of genes critical for biofilm establishment and maintenance [130].

Transcriptomic profiling also provides crucial information on mechanisms of drug resistance. Analysis of over a thousand *Plasmodium falciparum* isolates identified that upregulation of the unfolded protein response and slower progression through the early stages of the asexual intraerythrocytic developmental cycle were associated with artemisinin resistance in isolates from Southeast Asia [131].

Gene function annotation

All transcriptomic techniques have been particularly useful in identifying the functions of genes and identifying those responsible for particular phenotypes. Transcriptomics of *Arabidopsis* ecotypes that hyperaccumulate metals correlated genes involved in metal uptake, tolerance, and homeostasis with the phenotype [132]. Integration of RNA-Seq datasets across different tissues has been used to improve annotation of gene functions in commercially important organisms (e.g., cucumber) [133] or threatened species (e.g., koala) [134].

Assembly of RNA-Seq reads is not dependent on a reference genome [104], and it is so ideal for gene expression studies of nonmodel organisms with nonexisting or poorly developed genomic resources. For example, a database of SNPs used in Douglas fir breeding programs was created by de novo transcriptome analysis in the absence of a sequenced genome [135]. Similarly, genes that function in the development of cardiac, muscle, and nervous tissue in lobster were identified by comparing the transcriptomes of the various tissue types without use of a genome sequence [136]. RNA-Seq can also be used to identify previously unknown protein coding regions in existing sequenced genomes.

Noncoding RNA

Transcriptomics is most commonly applied to the mRNA content of the cell. However, the same techniques are equally applicable to noncoding RNAs that are not translated into a protein, but instead, have direct functions (e.g., roles in protein translation, DNA replication, RNA splicing, and transcriptional regulation) [137][138][139][140]. Many of these noncoding RNAs affect disease states, including cancer, cardiovascular, and neurological diseases [141].

Transcriptome databases

Transcriptomics studies generate large amounts of data that has potential applications far beyond the original aims of an experiment. As such, raw or processed data may be deposited

Table 5. Transcriptomic databases.

Name	Host	Data	Description
Gene Expression Omnibus [142]	NCBI	Microarray RNA-Seq	First transcriptomics database to accept data from any source. Introduced MIAME and MINSEQE community standards that define necessary experiment metadata to ensure effective interpretation and repeatability [143][144].
ArrayExpress [145]	ENA	Microarray	Imports datasets from the Gene Expression Omnibus and accepts direct submissions. Processed data and experiment metadata are stored at ArrayExpress, while the raw sequence reads are held at the ENA. Complies with MIAME and MINSEQE standards [144] [145].
Expression Atlas [146]	EBI	Microarray RNA-Seq	Tissue-specific gene expression database for animals and plants. Displays secondary analyses and visualisation, such as functional enrichment of Gene Ontology terms, InterPro domains, or pathways. Links to protein abundance data where available.
Genevestigator [147]	Privately curated	Microarray RNA-Seq	Contains manual curations of public transcriptome datasets, focusing on medical and plant biology data. Individual experiments are normalised across the full database, to allow comparison of gene expression across diverse experiments. Full functionality requires licence purchase, with free access to a limited functionality.
RefEx [148]	DDBJ	All	Human, mouse, and rat transcriptomes from 40 different organs. Gene expression visualised as heatmaps projected onto 3D representations of anatomical structures.
NONCODE [149]	noncode.org	RNA-Seq	ncRNAs excluding tRNA and rRNA.

DDBJ, DNA Data Bank of Japan; EBI, European Bioinformatics Institute; ENA, European Nucleotide Archive; MIAME, Minimum Information About a Microarray Experiment; MINSEQE, Minimum Information about a high-throughput nucleotide SEQuencing Experiment; NCBI, National Center for Biotechnology Information; ncRNAs, noncoding RNAs; RNA-Seq, RNA sequencing.

https://doi.org/10.1371/journal.pcbi.1005457.t005

into public databases to ensure their utility for the broader scientific community (Table 5). For example, as of 2016, the Gene Expression Omnibus contained millions of experiments.

Conclusions

Transcriptomics has revolutionised our understanding of how genomes are expressed. Over the last three decades, new technologies have redefined what is possible to investigate, and integration with other omics technologies is giving an increasingly integrated view of the complexities of cellular life. The plummeting cost of transcriptomics studies have made them possible for small laboratories, and large-scale transcriptomics consortia are able to undertake experiments comparing transcriptomes of thousands of organisms, tissues, or environmental conditions. This trend is likely to continue as sequencing technologies improve.

Supporting information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews.
(XML)

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991; 252:1651–6. PMID: 2047873
2. Pan Q, Shai O, Lee LJ, Frey BJ & Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 2008; 40:1413–5. https://doi.org/10.1038/ng.259 PMID: 18978789

3. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321:956–60. https://doi.org/10.1126/science.1160342 PMID: 18599741
4. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–11. https://doi.org/10.1038/nature12531 PMID: 24037378
5. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015; 348:660–5. https://doi.org/10.1126/science.aaa0355 PMID: 25954002
6. Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. Nat. Methods. 2014; 11:22–4. PMID: 24524133
7. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol. Cell. 2015; 58:610–20. https://doi.org/10.1016/j.molcel.2015.04.005 PMID: 26000846
8. McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol. 2013; 17:4–11. https://doi.org/10.1016/j.cbpa.2012.12.008 PMID: 23290152
9. Wang Z, Gerstein M & Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009; 10:57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660
10. Ozsolak F & Milos PM. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 2011; 12:87–98. https://doi.org/10.1038/nrg2934 PMID: 21191423
11. Morozova O, Hirst M & Marra MA. Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet. 2009; 10:135–51. https://doi.org/10.1146/annurev-genom-082908-145957 PMID: 19715439
12. Medline trend: automated yearly statistics of PubMed results for any query. [Internet]. Alexandru Dan Corlan [cited 2017 Apr 27]. http://dan.corlan.net/medline-trend.html.
13. Sim GK, Kafatos FC, Jones CW, Koehler MD, Efstratiadis A & Maniatis T. Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. Cell. 1979; 18:1303–16. PMID: 519770
14. Sutcliffe JG, Milner RJ, Bloom FE & Lerner RA. Common 82-nucleotide sequence unique to brain RNA. Proc. Natl. Acad. Sci. U.S.A. 1982; 79:4942–6. PMID: 6956902
15. Putney SD, Herlihy WC & Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature. 1983; 302:718–21. PMID: 6687628
16. Marra MA, Hillier L & Waterston RH. Expressed sequence tags—ESTablishing bridges between genomes. Trends Genet. 1998; 14:4–7. https://doi.org/10.1016/S0168-9525(97)01355-3 PMID: 9448457
17. Alwine JC, Kemp DJ & Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U.S.A. 1977; 74:5350–4. PMID: 414220
18. Becker-André M & Hahlbrock K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). Nucleic Acids Res. 1989; 17:9437–46. PMID: 2479917
19. Piétu G, Mariage-Samson R, Fayein NA, Matingou C, Eveno E, Houlgatte R, et al. The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. Genome Res. 1999; 9:195–209. PMID: 10022985
20. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, et al. Characterization of the yeast transcriptome. Cell. 1997; 88:243–51. PMID: 9008165
21. Velculescu VE, Zhang L, Vogelstein B & Kinzler KW. Serial analysis of gene expression. Science. 1995; 270:484–7. PMID: 7570003
22. Audic S & Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997; 7:986–95. PMID: 9331369
23. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res. 2014; 20:138–42. https://doi.org/10.12659/MSMBR.892101 PMID: 25149683
24. Zhao S, Fung-Leung WP, Bittner A, Ngo K & Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE. 2014; 9:e78644; https://doi.org/10.1371/journal.pone.0078644 PMID: 24454679

25. Hashimshony T, Wagner F, Sher N & Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012; 2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003 PMID: 22939981

26. Stears RL, Getts RC & Gullans SR. A novel, sensitive detection system for high-density microarrays using dendrimer technology. Physiol. Genomics. 2000; 3:93–9. PMID: 11015604

27. Illumina. RNA-Seq Data Comparison with Gene Expression Microarrays. European Pharmaceutical Review.

28. Black MB, Parks BB, Pluta L, Chu TM, Allen BC, Wolfinger RD, et al. Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments Toxicol. Sci. 2014; 137:385–403.

29. Marioni JC, Mason CE, Mane SM, Stephens M & Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18:1509–17. https://doi.org/10.1101/gr.079558.108 PMID: 18550803

30. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 2014; 32:903–14. https://doi.org/10.1038/nbt.2957 PMID: 25150838

31. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ & Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics. 2007; 8:412. https://doi.org/10.1186/1471-2105-8-412 PMID: 17961233

32. Larkin JE, Frank BC, Gavras H, Sultana R & Quackenbush J. Independence and reproducibility across microarray platforms. Nat. Methods. 2005; 2:337–44. https://doi.org/10.1038/nmeth757 PMID: 15846360

33. Nelson NJ. Microarrays have arrived: gene expression tool matures. J. Natl. Cancer Inst. 2001; 93:492–4. PMID: 11287436

34. Schena M, Shalon D, Davis RW & Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270:467–70. PMID: 7569999

35. Pozhitkov AE, Tautz D & Noble PA. Oligonucleotide microarrays: widely applied—poorly understood. Brief Funct Genomic Proteomic. 2007; 6:141–8. https://doi.org/10.1093/bfgp/elm014 PMID: 17644526

36. Heller MJ. DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng. 2002; 4:129–53. https://doi.org/10.1146/annurev.bioeng.4.020702.153438 PMID: 12117754

37. Ambroise, McLachlan Geoffrey J., Do Kim-Anh, Christopher. Analyzing Microarray Gene Expression Data. Hoboken: John Wiley & Sons. 2005;ISBN:9780471726128.

38. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays Nat. Biotechnol. 2000; 18:630–4. https://doi.org/10.1038/76469 PMID: 10835600

39. Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, et al. Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. Nat. Biotechnol. 2004; 22:1006–11. https://doi.org/10.1038/nbt992 PMID: 15247925

40. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics. 2006; 7:246. https://doi.org/10.1186/1471-2164-7-246 PMID: 17010196

41. Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5:621–8. https://doi.org/10.1038/nmeth.1226 PMID: 18516045

42. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453:1239–43. https://doi.org/10.1038/nature07002 PMID: 18488015

43. Chomczynski P & Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. 1987; 162:156–9. https://doi.org/10.1006/abio.1987.9999 PMID: 2440339

44. Chomczynski P & Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat Protoc. 2006; 1:581–5. https://doi.org/10.1038/nprot.2006.83 PMID: 17406285

45. Grillo M & Margolis FL. Use of reverse transcriptase polymerase chain reaction to monitor expression of intronless genes. BioTechniques. 1990; 9:262, 264, 266–8. PMID: 1699561

46. Bryant S & Manning DL. Isolation of messenger RNA. Methods Mol. Biol. 1998; 86:61–4. https://doi.org/10.1385/0-89603-494-1:61 PMID: 9664454

47. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN & Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014; 15:419. https://doi.org/10.1186/1471-2164-15-419 PMID: 24888378

48. Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, et al. A new resource for cereal genomics: 22K barley GeneChip comes of age. Plant Physiol. 2004; 134:960–8. https://doi.org/10.1104/pp.103.034462 PMID: 15020760

49. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U.S.A. 2003; 100:15776–81. https://doi.org/10.1073/pnas.2136655100 PMID: 14663149

50. Romanov V, Davidoff SN, Miles AR, Grainger DW, Gale BK & Brooks BD. A critical comparison of protein microarray fabrication technologies. Analyst. 2014; 139:1303–26. https://doi.org/10.1039/c3an01577g PMID: 24479125

51. Barbulovic-Nad I, Lucente M, Sun Y, Zhang M, Wheeler AR & Bussmann M. Bio-microarray fabrication techniques—a review. Crit. Rev. Biotechnol. 2006; 26:237–59. https://doi.org/10.1080/07388550600978358 PMID: 17095434

52. Auburn RP, Kreil DP, Meadows LA, Fischer B, Matilla SS & Russell S. Robotic spotting of cDNA and oligonucleotide microarrays. Trends Biotechnol. 2005; 23:374–9. https://doi.org/10.1016/j.tibtech.2005.04.002 PMID: 15978318

53. Shalon D, Smith SJ & Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 1996; 6:639–45. PMID: 8796352

54. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 1996; 14:1675–80. https://doi.org/10.1038/nbt1296-1675 PMID: 9634850

55. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B & Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003; 31:e15. PMID: 12582260

56. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, Brothman AR & Stallings RL. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer. 2005; 44:305–19. https://doi.org/10.1002/gcc.20243 PMID: 16075461

57. Tachibana Chris. Transcriptomics today: Microarrays, RNA-seq, and more. Science.

58. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320:1344–9. https://doi.org/10.1126/science.1158441 PMID: 18451266

59. Su Z, Fang H, Hong H, Shi L, Zhang W, Zhang W, et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. Genome Biol. 2014; 15:523. https://doi.org/10.1186/s13059-014-0523-y PMID: 25633159

60. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed subcellular RNA sequencing in situ. Science. 2014; 343:1360–3. https://doi.org/10.1126/science.1250212 PMID: 24578530

61. Shendure J & Ji H. Next-generation DNA sequencing. Nat. Biotechnol. 2008; 26:1135–45. https://doi.org/10.1038/nbt1486 PMID: 18846087

62. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014; 15:R86. https://doi.org/10.1186/gb-2014-15-6-r86 PMID: 24981968

63. Knierim E, Lucke B, Schwarz JM, Schuelke M & Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS ONE. 2011; 6:e28240; https://doi.org/10.1371/journal.pone.0028240 PMID: 22140562

64. Parekh S, Ziegenhain C, Vieth B, Enard W & Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016; 6:25533; https://doi.org/10.1038/srep25533 PMID: 27156886

65. Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. J Biomol Tech. 2015; 26:4–18. https://doi.org/10.7171/jbt.15-2601-001 PMID: 25649271

66. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011; 21:1543–51. https://doi.org/10.1101/gr.121095.111 PMID: 21816910

67. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods. 2011; 9:72–4. https://doi.org/10.1038/nmeth.1778 PMID: 22101854

68. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods. 2009; 6:377–82. https://doi.org/10.1038/nmeth.1315 PMID: 19349980

69. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods. 2014; 11:163–6. https://doi.org/10.1038/nmeth.2772 PMID: 24363023

70. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–9. https://doi.org/10.1126/science.1247651 PMID: 24531970

71. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods. 2010; 7:709–15. https://doi.org/10.1038/nmeth.1491 PMID: 20711195

72. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341. https://doi.org/10.1186/1471-2164-13-341 PMID: 22827831

73. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012; 2012:251364; https://doi.org/10.1155/2012/251364 PMID: 22829749

74. SRA. [Internet]. NCBI [cited 2017 April 27]. https://www.ncbi.nlm.nih.gov/sra.

75. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 2012; 30:434–9. https://doi.org/10.1038/nbt.2198 PMID: 22522955

76. Goodwin S, McPherson JD & McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 2016; 17:333–51. https://doi.org/10.1038/nrg.2016.49 PMID: 27184599

77. Garalde D, Snell E, Jachimowicz D, Heron A, Bruce Mark, Lloyd J, et al. Highly parallel direct RNA sequencing on an array of nanopores. bioRXiv. 2016.

78. Loman NJ, Quick J & Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods. 2015; 12:733–5. https://doi.org/10.1038/nmeth.3444 PMID: 26076426

79. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, et al. Direct RNA sequencing. Nature. 2009; 461:814–8. https://doi.org/10.1038/nature08390 PMID: 19776739

80. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013; 14:R95. https://doi.org/10.1186/gb-2013-14-9-r95 PMID: 24020486

81. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. https://doi.org/10.1038/nature11247 PMID: 22955616

82. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016; 44:D726–32. https://doi.org/10.1093/nar/gkv1160 PMID: 26527727

83. ENCODE: Encyclopedia of DNA Elements. [Internet]. ENCODE [cited 2017 Apr 27]. http://www.encodeproject.org.

84. Hart SN, Therneau TM, Zhang Y, Poland GA & Kocher JP. Calculating sample size estimates for RNA sequencing data. J. Comput. Biol. 2013; 20:970–8. https://doi.org/10.1089/cmb.2012.0283 PMID: 23961961

85. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401

86. Kodama Y, Shumway M & Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012; 40:D54–6. https://doi.org/10.1093/nar/gkr854 PMID: 22009675

87. Petrov A & Shams S. Microarray Image Processing and Quality Control. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology 38. 2004;(3): 211–226.

88. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011; 39:e90. https://doi.org/10.1093/nar/gkr344 PMID: 21576222

89. Van Verk MC, Hickman R, Pieterse CM & Van Wees SC. RNA-Seq: revelation of the messengers. Trends Plant Sci. 2013; 18:175–9. https://doi.org/10.1016/j.tplants.2013.02.001 PMID: 23481128

90. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods. 2015; 12:115–21. https://doi.org/10.1038/nmeth.3252 PMID: 25633503

91. FastQC: a quality control tool for high throughput sequence data. [Internet]. Babraham Institute [cited 2017 Apr 27]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

92. Lo CC & Chain PS. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics. 2014; 15:366. https://doi.org/10.1186/s12859-014-0366-2 PMID: 25408143

93. HTS Mappers. [Internet]. European Bioinformatics Institute [cited 2017 Apr 27]. http://www.ebi.ac.uk/~nf/hts_mappers/.

94. Fonseca NA, Rung J, Brazma A & Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012; 28:3169–77. https://doi.org/10.1093/bioinformatics/bts605 PMID: 23060614

95. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ & Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 2010; 28:511–5. https://doi.org/10.1038/nbt.1621 PMID: 20436464

96. Miller JR, Koren S & Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010; 95:315–27. https://doi.org/10.1016/j.ygeno.2010.03.001 PMID: 20211242

97. O'Neil ST & Emrich SJ. Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics. 2013; 14:465. https://doi.org/10.1186/1471-2164-14-465 PMID: 23837739

98. Smith-Unna R, Boursnell C, Patro R, Hibberd JM & Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016; 26:1134–44. https://doi.org/10.1101/gr.196469.115 PMID: 27252236

99. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R & Dewey CN. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014; 15:553. https://doi.org/10.1186/s13059-014-0553-5 PMID: 25608678

100. Zerbino DR & Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–9. https://doi.org/10.1101/gr.074492.107 PMID: 18349386

101. Schulz MH, Zerbino DR, Vingron M & Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012; 28:1086–92. https://doi.org/10.1093/bioinformatics/bts094 PMID: 22368243

102. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014; 30:1660–6. https://doi.org/10.1093/bioinformatics/btu077 PMID: 24532719

103. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat. Methods. 2010; 7:909–12. https://doi.org/10.1038/nmeth.1517 PMID: 20935650

104. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 2011; 29:644–52. https://doi.org/10.1038/nbt.1883 PMID: 21572440

105. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013; 8:1494–512. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962

106. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 2004; 14:1147–59. https://doi.org/10.1101/gr.1917404 PMID: 15140833

107. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–80. https://doi.org/10.1038/nature03959 PMID: 16056220

108. Kumar S & Blaxter ML. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics. 2010; 11:571. https://doi.org/10.1186/1471-2164-11-571 PMID: 20950480

109. Anders S, Pyl PT & Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–9. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700

110. Bray NL, Pimentel H, Melsted P & Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016; 34:525–7. https://doi.org/10.1038/nbt.3519 PMID: 27043002

111. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL & Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 2013; 31:46–53. https://doi.org/10.1038/nbt.2450 PMID: 23222703

112. Robinson MD, McCarthy DJ & Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

113. Love MI, Huber W & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

114. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

115. Fang Z & Cui X. Design and validation issues in RNA-seq experiments. Brief. Bioinformatics. 2011; 12:280–7. https://doi.org/10.1093/bib/bbr004 PMID: 21498551

116. Ramsköld D, Wang ET, Burge CB & Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 2009; 5:e1000598; https://doi.org/10.1371/journal.pcbi.1000598 PMID: 20011106

117. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002; 3:RESEARCH0034; PMID: 12184808

118. Core LJ, Waterfall JJ & Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–8. https://doi.org/10.1126/science.1162228 PMID: 19056941

119. Camarena L, Bruno V, Euskirchen G, Poggio S & Snyder M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. PLoS Pathog. 2010; 6:e1000834; https://doi.org/10.1371/journal.ppat.1000834 PMID: 20368969

120. Costa V, Aprile M, Esposito R & Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur. J. Hum. Genet. 2013; 21:134–42. https://doi.org/10.1038/ejhg.2012.129 PMID: 22739340

121. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA & Gerstein M. Role of non-coding sequence variants in cancer. Nat. Rev. Genet. 2016; 17:93–108. https://doi.org/10.1038/nrg.2015.17 PMID: 26781813

122. Slotkin RK & Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 2007; 8:272–85. https://doi.org/10.1038/nrg2072 PMID: 17363976

123. Proserpio V & Mahata B. Single-cell technologies to study the immune system. Immunology. 2016; 147:133–40. https://doi.org/10.1111/imm.12553 PMID: 26551575

124. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD & Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat. Rev. Genet. 2016; 17:257–71. https://doi.org/10.1038/nrg.2016.10 PMID: 26996076

125. Wu HJ, Wang AH & Jennings MP. Discovery of virulence factors of pathogenic bacteria. Curr Opin Chem Biol. 2008; 12:93–101. https://doi.org/10.1016/j.cbpa.2008.01.023 PMID: 18284925

126. Suzuki S, Horinouchi T & Furusawa C. Prediction of antibiotic resistance by gene expression profiles. Nat Commun. 2014; 5:5792; https://doi.org/10.1038/ncomms6792 PMID: 25517437

127. Westermann AJ, Gorski SA & Vogel J. Dual RNA-seq of pathogen and host. Nat. Rev. Microbiol. 2012; 10:618–30. https://doi.org/10.1038/nrmicro2852 PMID: 22890146

128. Durmuş S, Çakır T, Özgür A & Guthke R. A review on computational systems biology of pathogen-host interactions. Front Microbiol. 2015; 6:235. https://doi.org/10.3389/fmicb.2015.00235 PMID: 25914674

129. Garg R, Shankar R, Thakkar B, Kudapa H, Krishnamurthy L, Mantri N, et al. Transcriptome analyses reveal genotype- and developmental stage-specific molecular responses to drought and salinity stresses in chickpea. Sci Rep. 2016; 6:19228; https://doi.org/10.1038/srep19228 PMID: 26759178

130. García-Sánchez S, Aubert S, Iraqui I, Janbon G, Ghigo JM & d'Enfert C. Candida albicans biofilms: a developmental state associated with specific and stable gene expression patterns. Eukaryotic Cell. 2004; 3:536–45. https://doi.org/10.1128/EC.3.2.536-545.2004 PMID: 15075282

131. Mok S, Ashley EA, Ferreira PE, Zhu L, Lin Z, Yeo T, Chotivanich K, et al. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science. 2015; 347:431–5. https://doi.org/10.1126/science.1260403 PMID: 25502316

132. Verbruggen N, Hermans C & Schat H. Molecular mechanisms of metal hyperaccumulation in plants. New Phytol. 2009; 181:759–76. https://doi.org/10.1111/j.1469-8137.2008.02748.x PMID: 19192189

133. Li Z, Zhang Z, Yan P, Huang S, Fei Z & Lin K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genomics. 2011; 12:540. https://doi.org/10.1186/1471-2164-12-540 PMID: 22047402

134. Hobbs M, Pavasovic A, King AG, Prentis PJ, Eldridge MD, Chen Z, et al. A transcriptome resource for the koala (Phascolarctos cinereus): insights into koala retrovirus transcription and sequence diversity. BMC Genomics. 2014; 15:786. https://doi.org/10.1186/1471-2164-15-786 PMID: 25214207

135. Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dolan P, et al. A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. BMC Genomics. 2013; 14:137. https://doi.org/10.1186/1471-2164-14-137 PMID: 23445355

136. McGrath LL, Vollmer SV, Kaluziak ST & Ayers J. De novo transcriptome assembly for the lobster Homarus americanus and characterization of differential gene expression across nervous system tissues. BMC Genomics. 2016; 17:63. https://doi.org/10.1186/s12864-016-2373-3 PMID: 26772543

137. Noller HF. Ribosomal RNA and translation. Annu. Rev. Biochem. 1991; 60:191–227. https://doi.org/10.1146/annurev.bi.60.070191.001203 PMID: 1883196

138. Christov CP, Gardiner TJ, Szüts D & Krude T. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol. Cell. Biol. 2006; 26:6993–7004; https://doi.org/10.1128/MCB.01060-06 PMID: 16943439

139. Kishore S & Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science. 2006; 311:230–2. https://doi.org/10.1126/science.1118265 PMID: 16357227

140. Hüttenhofer A, Schattner P & Polacek N. Non-coding RNAs: hope or hype?. Trends Genet. 2005; 21:289–97. https://doi.org/10.1016/j.tig.2005.03.007 PMID: 15851066

141. Esteller M. Non-coding RNAs in human disease. Nat. Rev. Genet. 2011; 12:861–74. https://doi.org/10.1038/nrg3074 PMID: 22094949

142. Edgar R, Domrachev M & Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207–10. PMID: 11752295

143. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. 2001; 29:365–71. https://doi.org/10.1038/ng1201-365 PMID: 11726920

144. Brazma A. Minimum Information About a Microarray Experiment (MIAME)—successes, failures, challenges. ScientificWorldJournal. 2009; 9:420–3. https://doi.org/10.1100/tsw.2009.57 PMID: 19484163

145. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015; 43:D1113–6. https://doi.org/10.1093/nar/gku1057 PMID: 25361974

146. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 2016; 44:D746–52. https://doi.org/10.1093/nar/gkv1045 PMID: 26481351

147. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, et al. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics. 2008; 2008:420747; https://doi.org/10.1155/2008/420747 PMID: 19956698

148. Mitsuhashi N, Fujieda K, Tamura T, Kawamoto S, Takagi T & Okubo K. BodyParts3D: 3D structure database for anatomical concepts. Nucleic Acids Res. 2009; 37:D782–5. https://doi.org/10.1093/nar/gkn613 PMID: 18835852

149. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, et al. NONCODE 2016;an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 2016; 44:D203–8. https://doi.org/10.1093/nar/gkv1252 PMID: 26586799

Check for updates

SOFTWARE TOOL ARTICLE

REVISED RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; referees: 3 approved]

Charity W. Law[1,2], Monther Alhamdoosh[3], Shian Su[1], Xueyi Dong[1], Luyi Tian[1,2], Gordon K. Smyth[1,4], Matthew E. Ritchie[1,2,4]

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Australia
[2]Department of Medical Biology, The University of Melbourne, Parkville, 3010, Australia
[3]CSL Limited, Parkville, Victoria, 3010, Australia
[4]School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, Australia

v3 **First published:** 17 Jun 2016, **5**:1408 (https://doi.org/10.12688/f1000research.9005.1)
Second version: 30 Nov 2016, **5**:1408 (https://doi.org/10.12688/f1000research.9005.2)
Latest published: 28 Dec 2018, **5**:1408 (https://doi.org/10.12688/f1000research.9005.3)

Abstract
The ability to easily and efficiently analyse RNA-sequencing data is a key strength of the Bioconductor project. Starting with counts summarised at the gene-level, a typical analysis involves pre-processing, exploratory data analysis, differential expression testing and pathway analysis with the results obtained informing future experiments and validation studies. In this workflow article, we analyse RNA-sequencing data from the mouse mammary gland, demonstrating use of the popular **edgeR** package to import, organise, filter and normalise the data, followed by the **limma** package with its *voom* method, linear modelling and empirical Bayes moderation to assess differential expression and perform gene set testing. This pipeline is further enhanced by the **Glimma** package which enables interactive exploration of the results so that individual samples and genes can be examined by the user. The complete analysis offered by these three packages highlights the ease with which researchers can turn the raw counts from an RNA-sequencing experiment into biological insights using Bioconductor.

Keywords
RNA sequencing, data analysis, gene expression

This article is included in the International Society for Computational Biology Community Journal gateway.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 3 published 28 Dec 2018			
REVISED version 2 published 30 Nov 2016			
version 1 published 17 Jun 2016	✓ report	✓ report	✓ report

1 **James W. MacDonald**, University of Washington, USA

2 **Maria A. Doyle**, Peter MacCallum Cancer Centre, Australia

3 **Jovana Maksimovic**, Murdoch Children's Research Institute, Australia

Discuss this article

Comments (1)



This article is included in the Bioconductor gateway.

Corresponding authors: Gordon K. Smyth (smyth@wehi.edu.au), Matthew E. Ritchie (mritchie@wehi.edu.au)

Author roles: Law CW: Conceptualization, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Alhamdoosh M**: Software, Writing – Review & Editing; **Su S**: Software, Writing – Review & Editing; **Dong X**: Software, Writing – Review & Editing; **Tian L**: Software, Writing – Review & Editing; **Smyth GK**: Funding Acquisition, Software, Supervision, Writing – Review & Editing; **Ritchie ME**: Conceptualization, Funding Acquisition, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the National Health and Medical Research Council (NHMRC) (Fellowship GNT1058892 and Program GNT1054618 to GKS, Project GNT1050661 to MER and GKS and Fellowship GNT1104924 to MER), Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Law CW *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Law CW, Alhamdoosh M, Su S *et al.* **RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; referees: 3 approved]** *F1000Research* 2018, **5**:1408 (https://doi.org/10.12688/f1000research.9005.3)

First published: 17 Jun 2016, **5**:1408 (https://doi.org/10.12688/f1000research.9005.1)

REVISED **Amendments from Version 2**

The updated workflow makes use of current versions of software: R version 3.5.1 and Bioconductor project version 3.8. The filtering strategy has been relaxed, using default settings in the filterByExpr function from the edgeR package which retains approximately 2500 more genes than the previous version. Output downstream of filtering has been updated, including adjustment to the vertical dotted line in Figure 1 marking the new log-CPM threshold. In "Transformations from the raw-scale" and "Removing genes that are lowly expressed", text has been added to give more details on log-CPM values that are calculated and gene filtering strategy. Glimma MD plot and heatmap now uses lcpm values to represent expression. The reference for Glimma has been updated, and the id.column argument in the glMDPlot function was renamed to side.main. Placement of some figures have been adjusted so that they appear around its associated text, which previously affected the pdf version of the article. Xueyi Dong and Luyi Tian are added as authors for translation of the article to Chinese which is available in the release version of the RNAseq123 workflow package from Bioconductor, http://bioconductor.org/packages/RNAseq123. Xueyi Dong also updated the workflow to Bioconductor 3.8. These changes have also been outlined in "Software availability" and "Author contributions".

See referee reports

Introduction

RNA-sequencing (RNA-seq) has become the primary technology used for gene expression profiling, with the genome-wide detection of differentially expressed genes between two or more conditions of interest one of the most commonly asked questions by researchers. The **edgeR**[1] and **limma** packages[2] available from the Bioconductor project[3] offer a well-developed suite of statistical methods for dealing with this question for RNA-seq data. In this article, we describe an **edgeR** - **limma** workflow for analysing RNA-seq data that takes gene-level counts as its input, and moves through pre-processing and exploratory data analysis before obtaining lists of differentially expressed (DE) genes and gene signatures. This analysis is enhanced through the use of interactive graphics from the **Glimma** package[4], that allows for a more detailed exploration of the data at both the sample and gene-level than is possible using static `R` plots.

The experiment analysed in this workflow is from Sheridan *et al.* (2015)[5] and consists of three cell populations (basal, luminal progenitor (LP) and mature luminal (ML)) sorted from the mammary glands of female virgin mice, each profiled in triplicate. RNA samples were sequenced across three batches on an Illumina HiSeq 2000 to obtain 100 base-pair single-end reads. The analysis outlined in this article assumes that reads obtained from an RNA-seq experiment have been aligned to an appropriate reference genome and summarised into counts associated with gene-specific regions. In this instance, reads were aligned to the mouse reference genome (mm10) using the `R` based pipeline available in the **Rsubread** package (specifically the `align` function[6] followed by `featureCounts`[7] for gene-level summarisation based on the in-built *mm10* RefSeq-based annotation). Count data for these samples can be downloaded from the Gene Expression Omnibus (GEO) `http://www.ncbi.nlm.nih.gov/geo/` using GEO Series accession number GSE63310. Further information on experimental design and sample preparation is also available from GEO under this accession number.

Data packaging

Reading in count data

To get started with this analysis, download the file `GSE63310_RAW.tar` available online from `http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE63310&format=file`, and extract the relevant files from this archive. Each of these text files contains the raw gene-level counts for a given sample. Note that our analysis only includes the basal, LP and ML samples from this experiment (see associated file names below).

```
files <- c("GSM1545535_10_6_5_11.txt", "GSM1545536_9_6_5_11.txt",
   "GSM1545538_purep53.txt", "GSM1545539_JMS8-2.txt",
   "GSM1545540_JMS8-3.txt", "GSM1545541_JMS8-4.txt",
   "GSM1545542_JMS8-5.txt", "GSM1545544_JMS9-P7c.txt",
   "GSM1545545_JMS9-P8c.txt")
read.delim(files[1], nrow=5)
```

```
##    EntrezID GeneLength Count
## 1    497097       3634     1
## 2 100503874       3259     0
## 3 100038431       1634     0
```

```
## 4      19888        9747      0
## 5      20671        3130      1
```

Whilst each of the nine text files can be read into R separately and combined into a matrix of counts, **edgeR** offers a convenient way to do this in one step using the readDGE function. The resulting DGEList-object contains a matrix of counts with 27,179 rows associated with unique Entrez gene identifiers (IDs) and nine columns associated with the individual samples in the experiment.

```
library(limma)
library(edgeR)
x <- readDGE(files, columns=c(1,3))
class(x)
```

```
## [1] "DGEList"
## attr(,"package")
## [1] "edgeR"
```

```
dim(x)
```

```
## [1] 27179     9
```

If the counts from all samples were stored in a single file, the data can be read into R and then converted into a DGE-List-object using the DGEList function.

Organising sample information

For downstream analysis, sample-level information related to the experimental design needs to be associated with the columns of the counts matrix. This should include experimental variables, both biological and technical, that could have an effect on expression levels. Examples include cell type (basal, LP and ML in this experiment), genotype (wild-type, knock-out), phenotype (disease status, sex, age), sample treatment (drug, control) and batch information (date experiment was performed if samples were collected and analysed at distinct time points) to name just a few.

Our DGEList-object contains a samples data frame that stores both cell type (or group) and batch (sequencing lane) information, each of which consists of three distinct levels. Note that within x$samples, library sizes are automatically calculated for each sample and normalisation factors are set to 1. For simplicity, we remove the GEO sample IDs (GSM*) from the column names of our DGEList-object x.

```
samplenames <- substring(colnames(x), 12, nchar(colnames(x)))
samplenames
```

```
## [1] "10_6_5_11" "9_6_5_11"  "purep53"   "JMS8-2"    "JMS8-3"    "JMS8-4"    "JMS8-5"
## [8] "JMS9-P7c"  "JMS9-P8c"
```

```
colnames(x) <- samplenames
group <- as.factor(c("LP", "ML", "Basal", "Basal", "ML", "LP", "Basal", "ML", "LP"))
x$samples$group <- group
lane <- as.factor(rep(c("L004","L006","L008"), c(3,4,2)))
x$samples$lane <- lane
x$samples
```

```
##                              files group lib.size norm.factors lane
## 10_6_5_11 GSM1545535_10_6_5_11.txt    LP 32863052            1 L004
## 9_6_5_11   GSM1545536_9_6_5_11.txt    ML 35335491            1 L004
## purep53     GSM1545538_purep53.txt Basal 57160817            1 L004
## JMS8-2       GSM1545539_JMS8-2.txt Basal 51368625            1 L006
## JMS8-3       GSM1545540_JMS8-3.txt    ML 75795034            1 L006
## JMS8-4       GSM1545541_JMS8-4.txt    LP 60517657            1 L006
## JMS8-5       GSM1545542_JMS8-5.txt Basal 55086324            1 L006
```

```
## JMS9-P7c   GSM1545544_JMS9-P7c.txt    ML 21311068            1 L008
## JMS9-P8c   GSM1545545_JMS9-P8c.txt    LP 19958838            1 L008
```

Organising gene annotations

A second data frame named `genes` in the DGEList-object is used to store gene-level information associated with rows of the counts matrix. This information can be retrieved using organism specific packages such as **Mus.musculus**[8] for mouse (or **Homo.sapiens**[9] for human) or the **biomaRt** package[10,11] which interfaces the Ensembl genome databases in order to perform gene annotation. The type of information that can be retrieved includes gene symbols, gene names, chromosome names and locations, Entrez gene IDs, Refseq gene IDs and Ensembl gene IDs to name just a few. **biomaRt** primarily works off Ensembl gene IDs, whereas **Mus.musculus** packages information from various sources and allows users to choose between many different gene IDs as the key. The Entrez gene IDs available in our dataset were annotated using the **Mus.musculus** package to retrieve associated gene symbols and chromosome information.

```
library(Mus.musculus)
geneid <- rownames(x)
genes <- select(Mus.musculus, keys=geneid, columns=c("SYMBOL", "TXCHROM"),
                keytype="ENTREZID")
dim(genes)
```

```
## [1] 27220     3
```

```
head(genes)
```

```
##      ENTREZID   SYMBOL   TXCHROM
## 1      497097     Xkr4      chr1
## 2   100503874  Gm19938      <NA>
## 3   100038431  Gm10568      <NA>
## 4       19888      Rp1      chr1
## 5       20671    Sox17      chr1
## 6       27395   Mrpl15      chr1
```

As with any gene ID, Entrez gene IDs may not map one-to-one to the gene information of interest. It is important to check for duplicated gene IDs and to understand the source of duplication before resolving them. Our gene annotation contains 28 genes that map to multiple chromosomes (e.g. gene Gm1987 is associated with "chr4" and "chr4_JH584294_random" and microRNA Mir5098 is associated with "chr2", "chr5", "chr8", "chr11" and "chr17"). To resolve duplicate gene IDs one could combine all chromosome information from the multi-mapped genes, such that gene Gm1987 would be is assigned to "chr4 and chr4_JH584294_random", or select one of the chromosomes to represent the gene with duplicate annotation. For simplicity we do the latter, keeping only the first occurrence of each gene ID.

```
genes <- genes[!duplicated(genes$ENTREZID),]
```

In this example, the gene order is the same in both the annotation and the data object. If this is not the case due to missing and/or rearranged gene IDs, the `match` function can be used to order genes correctly. The data frame of gene annotations is then added to the data object and neatly packaged in a DGEList-object containing raw count data with associated sample information and gene annotations.

```
x$genes <- genes
x
```

```
## An object of class "DGEList"
## $samples
##                              files group lib.size norm.factors lane
## 10_6_5_11 GSM1545535_10_6_5_11.txt    LP 32863052            1 L004
## 9_6_5_11   GSM1545536_9_6_5_11.txt    ML 35335491            1 L004
## purep53     GSM1545538_purep53.txt Basal 57160817            1 L004
## JMS8-2       GSM1545539_JMS8-2.txt Basal 51368625            1 L006
```

```
## JMS8-3       GSM1545540_JMS8-3.txt    ML 75795034            1 L006
## JMS8-4       GSM1545541_JMS8-4.txt    LP 60517657            1 L006
## JMS8-5       GSM1545542_JMS8-5.txt Basal 55086324            1 L006
## JMS9-P7c   GSM1545544_JMS9-P7c.txt    ML 21311068            1 L008
## JMS9-P8c   GSM1545545_JMS9-P8c.txt    LP 19958838            1 L008
##
## $counts
##             Samples
## Tags         10_6_5_11 9_6_5_11 purep53 JMS8-2 JMS8-3 JMS8-4 JMS8-5 JMS9-P7c JMS9-P8c
##   497097             1       2     342    526      3      3    535        2        0
##   100503874          0       0       5      6      0      0      5        0        0
##   100038431          0       0       0      0      0      0      1        0        0
##   19888              0       1       0      0     17      2      0        1        0
##   20671              1       1      76     40     33     14     98       18        8
## 27174 more rows ...
##
## $genes
##    ENTREZID  SYMBOL  TXCHROM
## 1    497097    Xkr4     chr1
## 2 100503874 Gm19938     <NA>
## 3 100038431 Gm10568     <NA>
## 4     19888     Rp1     chr1
## 5     20671   Sox17     chr1
## 27174 more rows ...
```

Data pre-processing

Transformations from the raw-scale

For differential expression and related analyses, gene expression is rarely considered at the level of raw counts since libraries sequenced at a greater depth will result in higher counts. Rather, it is common practice to transform raw counts onto a scale that accounts for such library size differences. Popular transformations include counts per million (CPM), $\log_2$-counts per million (log-CPM), reads per kilobase of transcript per million (RPKM), and fragments per kilobase of transcript per million (FPKM).

In our analyses, CPM and log-CPM transformations are used regularly although they do not account for gene length differences as RPKM and FPKM values do. Whilst RPKM and FPKM values can just as well be used, CPM and log-CPM values can be calculated using a counts matrix alone and will suffice for the type of comparisons we are interested in. Assuming that there are no differences in isoform usage between conditions, differential expression analyses look at gene expression changes between conditions rather than comparing expression across multiple genes or drawing conclusions on absolute levels of expression. In other words, gene lengths remain constant for comparisons of interest and any observed differences are a result of changes in condition rather than changes in gene length.

Here raw counts are converted to CPM and log-CPM values using the `cpm` function in **edgeR**. RPKM values are just as easily calculated as CPM values using the `rpkm` function in **edgeR** if gene lengths are available.

```
cpm <- cpm(x)
lcpm <- cpm(x, log=TRUE)
```

A CPM value of 1 for a gene equates to having 20 counts in the sample with the lowest sequencing depth (JMS9-P8c, library size ≈20 million) or 76 counts in the sample with the greatest sequencing depth (JMS8-3, library size ≈76 million).

The log-CPM values will be used for exploratory plots. When `log=TRUE`, the `cpm` function adds an offset to the CPM values before converting to the log2-scale. By default, the offset is $2/L$ where 2 is the "prior count" and L is the average library size in millions, so the log-CPM values are related to the CPM values by $\log_2(\text{CPM} + 2/L)$. This calculation ensures that any two read counts with identical CPM values will also have identical log-CPM values. The prior count avoids taking the logarithm of zero, and also reduces spurious variability for genes

with very low counts by shrinking all the inter-sample log-fold-changes towards zero, something that is helpful for exploratory plotting. For this dataset, the average library size is about 45.5 million, so $L \approx 45.5$ and the minimum log-CPM value for each sample becomes $\log_2(2/45.5) = -4.51$. In other words, a count of zero for this data maps to a log-CPM value of −4.51 after adding the prior count or offset:

```
L <- mean(x$samples$lib.size) * 1e-6
M <- median(x$samples$lib.size) * 1e-6
c(L, M)
```

```
## [1] 45.5 51.4
```

```
summary(lcpm)
```

```
##   10_6_5_11        9_6_5_11        purep53          JMS8-2          JMS8-3
## Min.   :-4.51   Min.   :-4.51   Min.   :-4.51   Min.   :-4.51   Min.   :-4.51
## 1st Qu.:-4.51   1st Qu.:-4.51   1st Qu.:-4.51   1st Qu.:-4.51   1st Qu.:-4.51
## Median :-0.68   Median :-0.36   Median :-0.10   Median :-0.09   Median :-0.43
## Mean   : 0.17   Mean   : 0.33   Mean   : 0.44   Mean   : 0.41   Mean   : 0.32
## 3rd Qu.: 4.29   3rd Qu.: 4.56   3rd Qu.: 4.60   3rd Qu.: 4.55   3rd Qu.: 4.58
## Max.   :14.76   Max.   :13.50   Max.   :12.96   Max.   :12.85   Max.   :12.96
##     JMS8-4          JMS8-5          JMS9-P7c        JMS9-P8c
## Min.   :-4.51   Min.   :-4.51   Min.   :-4.51   Min.   :-4.51
## 1st Qu.:-4.51   1st Qu.:-4.51   1st Qu.:-4.51   1st Qu.:-4.51
## Median :-0.41   Median :-0.07   Median :-0.17   Median :-0.33
## Mean   : 0.25   Mean   : 0.40   Mean   : 0.37   Mean   : 0.27
## 3rd Qu.: 4.32   3rd Qu.: 4.43   3rd Qu.: 4.60   3rd Qu.: 4.44
## Max.   :14.85   Max.   :13.19   Max.   :12.94   Max.   :14.01
```

Log-CPM values are also used in downstream linear modeling via **limma**'s `voom` function, although `voom` recomputes its own log-CPM values internally with a smaller prior count.

Removing genes that are lowly expressed

All datasets will include a mix of genes that are expressed and those that are not expressed. Whilst it is of interest to examine genes that are expressed in one condition but not in another, some genes are unexpressed throughout all samples. In fact, 19% of genes in this dataset have zero counts across all nine samples.

```
table(rowSums(x$counts==0)==9)
```

```
##
## FALSE  TRUE
## 22026  5153
```

Plotting the distribution log-CPM values shows that a sizeable proportion of genes within each sample are either unexpressed or lowly-expressed with log-CPM values that are small or negative (Figure 1A).

Genes that do not have a worthwhile number of reads in any sample should be filtered out of the downstream analyses. There are several reasons for this. From a biological point of view, genes that not expressed at a biologically meaningful level in any condition are not of interest and are therefore best ignored. From a statistical point of view, removing low count genes allows the mean-variance relationship in the data to be estimated with greater reliability and also reduces the number of statistical tests that need to be carried out in downstream analyses looking at differential expression.

The `filterByExpr` function in the **edgeR** package provides an automatic way to filter genes, while keeping as many genes as possible with worthwhile counts.

```
keep.exprs <- filterByExpr(x, group=group)
x <- x[keep.exprs,, keep.lib.sizes=FALSE]
dim(x)
```

```
## [1] 16624     9
```

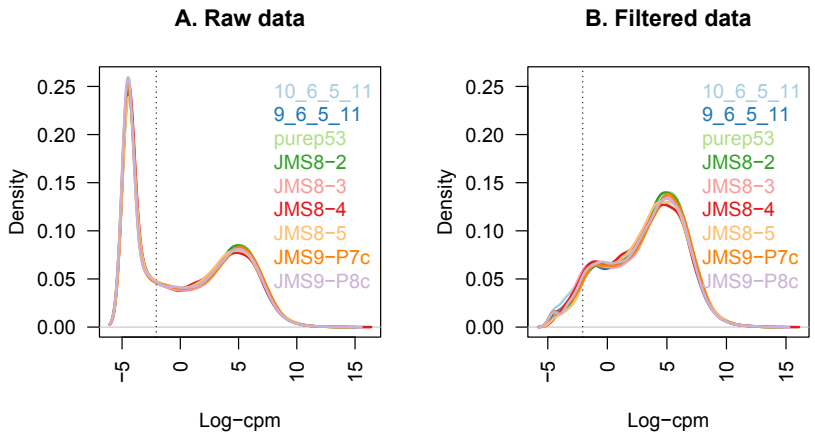


Figure 1. The density of log-CPM values for raw pre-filtered data (**A**) and post-filtered data (**B**) are shown for each sample. Dotted vertical lines mark the log-CPM threshold (equivalent to a CPM value of about 0.2) used in the filtering step.

By default, the function keeps genes with about 10 read counts or more in a minimum number of samples, where the number of samples is chosen according to the minimum group sample size. The actual filtering uses CPM values rather than counts in order to avoid giving preference to samples with large library sizes. For this dataset, the median library size is about 51 million and $10/51 \approx 0.2$, so the `filterByExpr` function keeps genes that have a CPM of 0.2 or more in at least three samples. A biologically interesting gene should be expressed in at least three samples because all the cell type groups have three replicates. The cutoffs used depend on the sequencing depth and on the experimental design. If the library sizes had been larger then a lower CPM cutoff would have been chosen, because larger library sizes provide better resolution to explore more genes at lower expression levels. Alternatively, smaller library sizes decrease our ability to explore marginal genes and hence would have led to a higher CPM cutoff.

Using this criterion, the number of genes is reduced to 16,624, about 60% of the number that we started with. The log-CPM values after filtering show a nearly unimodel distribution for each sample (Figure 1B). Note that subsetting the entire DGEList-object removes both the counts and the associated gene information for the filtered genes. The filtered DGEList-object keeps the gene information and the counts for the retained genes correctly associated.

Code to produce Figure 1 is given below:

```
lcpm.cutoff <- log2(10/M + 2/L)
library(RColorBrewer)
nsamples <- ncol(x)
col <- brewer.pal(nsamples, "Paired")
par(mfrow=c(1,2))
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="", xlab="")
title(main="A. Raw data", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)
for (i in 2:nsamples){
  den <- density(lcpm[,i])
  lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")
lcpm <- cpm(x, log=TRUE)
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="", xlab="")
title(main="B. Filtered data", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)
for (i in 2:nsamples){
  den <- density(lcpm[,i])
  lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")
```

Normalising gene expression distributions

During the sample preparation or sequencing process, external factors that are not of biological interest can affect the expression of individual samples. For example, samples processed in the first batch of an experiment can have higher expression overall when compared to samples processed in a second batch. It is assumed that all samples should have a similar range and distribution of expression values. Normalisation is required to ensure that the expression distributions of each sample are similar across the entire experiment.

Any plot showing the per sample expression distributions, such as a density or boxplot, is useful in determining whether any samples are dissimilar to others. Distributions of log-CPM values are similar throughout all samples within this dataset (Figure 1B).

Nonetheless, normalisation by the method of trimmed mean of M-values (TMM)[12] is performed using the `calcNormFactors` function in **edgeR**. The normalisation factors calculated here are used as a scaling factor for the library sizes. When working with DGEList-objects, these normalisation factors are automatically stored in `x$samples$norm.factors`. For this dataset the effect of TMM-normalisation is mild, as evident in the magnitude of the scaling factors, which are all relatively close to 1.

```
x <- calcNormFactors(x, method = "TMM")
x$samples$norm.factors
```

```
## [1] 0.894 1.025 1.046 1.046 1.016 0.922 0.996 1.086 0.984
```

To give a better visual representation of the effects of normalisation, the data was duplicated then adjusted so that the counts of the first sample are reduced to 5% of their original values, and in the second sample they are inflated to be 5-times larger.

```
x2 <- x
x2$samples$norm.factors <- 1
x2$counts[,1] <- ceiling(x2$counts[,1]*0.05)
x2$counts[,2] <- x2$counts[,2]*5
```

Figure 2 shows the expression distribution of samples for unnormalised and normalised data, where distributions are noticeably different pre-normalisation and are similar post-normalisation. Here the first sample has a small

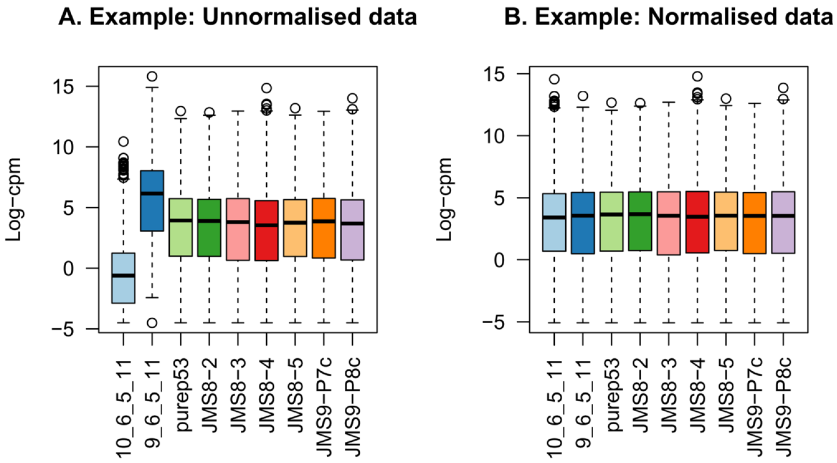


Figure 2. Example data: Boxplots of log-CPM values showing expression distributions for unnormalised data (**A**) and normalised data (**B**) for each sample in the modified dataset where the counts in samples 1 and 2 have been scaled to 5% and 500% of their original values respectively.

TMM scaling factor of 0.06, whereas the second sample has a large scaling factor of 6.08 – neither values are close to 1.

```
par(mfrow=c(1,2))
lcpm <- cpm(x2, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
title(main="A. Example: Unnormalised data", ylab="Log-cpm")
x2 <- calcNormFactors(x2)
x2$samples$norm.factors
```

```
## [1] 0.0577 6.0829 1.2202 1.1648 1.1966 1.0466 1.1505 1.2543 1.1090
```

```
lcpm <- cpm(x2, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
title(main="B. Example: Normalised data", ylab="Log-cpm")
```

Unsupervised clustering of samples

In our opinion, one of the most important exploratory plots to examine for gene expression analyses is the multi-dimensional scaling (MDS) plot, or similar. The plot shows similarities and dissimilarities between samples in an unsupervised manner so that one can have an idea of the extent to which differential expression can be detected before carrying out formal tests. Ideally, samples would cluster well within the primary condition of interest, and any sample straying far from its group could be identified and followed up for sources of error or extra variation. If present, technical replicates should lie very close to one another.

Such a plot can be made in **limma** using the `plotMDS` function. The first dimension represents the leading-fold-change that best separates samples and explains the largest proportion of variation in the data, with subsequent dimensions having a smaller effect and being orthogonal to the ones before it. When experimental design involves multiple factors, it is recommended that each factor is examined over several dimensions. If samples cluster by a given factor in any of these dimensions, it suggests that the factor contributes to expression differences and is worth including in the linear modelling. On the other hand, factors that show little or no effect may be left out of downstream analysis.

In this dataset, samples can be seen to cluster well within experimental groups over dimension 1 and 2, and then separate by sequencing lane (sample batch) over dimension 3 (Figure 3). Keeping in mind that the first dimension explains the largest proportion of variation in the data, notice that the range of values over the dimensions become

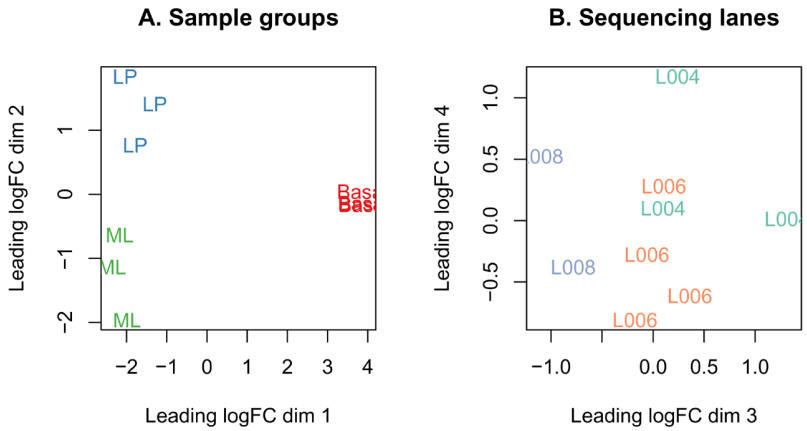


Figure 3. MDS plots of log-CPM values over dimensions 1 and 2 with samples coloured and labeled by sample groups (**A**) and over dimensions 3 and 4 with samples coloured and labeled by sequencing lane (**B**). Distances on the plot correspond to the leading fold-change, which is the average (root-mean-square) $\log_2$-fold-change for the 500 genes most divergent between each pair of samples by default.

smaller as we move to higher dimensions. Whilst all samples cluster by groups, the largest transcriptional difference is observed between basal and LP, and basal and ML over dimension 1. For this reason, it is expected that pairwise comparisons between cell populations will result in a greater number of DE genes for comparisons involving basal samples, and relatively small numbers of DE genes when comparing ML to LP. Datasets where samples do not cluster by experimental group may show little or no evidence of differential expression in the downstream analysis.

To create the MDS plots, we assign different colours to the factors of interest. Dimensions 1 and 2 are examined using the colour grouping defined by cell types.

```
lcpm <- cpm(x, log=TRUE)
par(mfrow=c(1,2))
col.group <- group
levels(col.group) <-  brewer.pal(nlevels(col.group), "Set1")
col.group <- as.character(col.group)
col.lane <- lane
levels(col.lane) <-  brewer.pal(nlevels(col.lane), "Set2")
col.lane <- as.character(col.lane)
plotMDS(lcpm, labels=group, col=col.group)
title(main="A. Sample groups")
```

Dimensions 3 and 4 are examined using the colour grouping defined by sequencing lanes (batch).

```
plotMDS(lcpm, labels=lane, col=col.lane, dim=c(3,4))
title(main="B. Sequencing lanes")
```

Alternatively, the **Glimma** package offers the convenience of an interactive MDS plot where multiple dimensions can be explored. The `glMDSPlot` function generates an html page (that is opened in a browser if `launch=TRUE`) with an MDS plot in the left panel and a barplot showing the proportion of variation explained by each dimension in the right panel. Clicking on the bars of the bar plot changes the pair of dimensions plotted in the MDS plot, and hovering over the individual points reveals the sample label. The colour scheme can be changed as well to highlight cell population or sequencing lane (batch). An interactive MDS plot of this dataset can be found at http://bioinf.wehi.edu.au/folders/limmaWorkflow/glimma-plots/MDS-Plot.html

```
library(Glimma)
glMDSPlot(lcpm, labels=paste(group, lane, sep="_"), groups=x$samples[,c(2,5)],
          launch=FALSE)
```

Differential expression analysis

Creating a design matrix and contrasts

In this study, it is of interest to see which genes are expressed at different levels between the three cell populations profiled. In our analysis, linear models are fitted to the data with the assumption that the underlying data is normally distributed. To get started, a design matrix is set up with both the cell population and sequencing lane (batch) information.

```
design <- model.matrix(~0+group+lane)
colnames(design) <- gsub("group", "", colnames(design))
design
```

```
##   Basal LP ML laneL006 laneL008
## 1     0  1  0        0        0
## 2     0  0  1        0        0
## 3     1  0  0        0        0
## 4     1  0  0        1        0
## 5     0  0  1        1        0
## 6     0  1  0        1        0
## 7     1  0  0        1        0
## 8     0  0  1        0        1
## 9     0  1  0        0        1
```

```
## attr(,"assign")
## [1] 1 1 1 2 2
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
##
## attr(,"contrasts")$lane
## [1] "contr.treatment"
```

For a given experiment, there are usually several equivalent ways to set up an appropriate design matrix. For example, `~0+group+lane` removes the intercept from the first factor, `group`, but an intercept remains in the second factor `lane`. Alternatively, `~group+lane` could be used to keep the intercepts in both `group` and `lane`. Understanding how to interpret the coefficients estimated in a given model is key here. We choose the first model for our analysis, as setting up model contrasts is more straight forward in the absence of an intercept for `group`. Contrasts for pairwise comparisons between cell populations are set up in **limma** using the `makeContrasts` function.

```
contr.matrix <- makeContrasts(
   BasalvsLP = Basal-LP,
   BasalvsML = Basal - ML,
   LPvsML = LP - ML,
   levels = colnames(design))
contr.matrix
```

```
##           Contrasts
## Levels     BasalvsLP BasalvsML LPvsML
##   Basal            1         1      0
##   LP              -1         0      1
##   ML               0        -1     -1
##   laneL006         0         0      0
##   laneL008         0         0      0
```

A key strength of **limma**'s linear modelling approach is the ability accommodate arbitrary experimental complexity. Simple designs, such as the one in this workflow, with cell type and batch, through to more complicated factorial designs and models with interaction terms can be handled relatively easily. Where experimental or technical effects can be modelled using a random effect, another possibility in **limma** is to estimate correlations using `duplicateCorrelation` by specifying a `block` argument for both this function and in the `lmFit` linear modelling step.

Removing heteroscedascity from count data

It has been shown that for RNA-seq count data, the variance is not independent of the mean[13] – this is true of raw counts or when transformed to log-CPM values. Methods that model counts using a Negative Binomial distribution assume a quadratic mean-variance relationship. In **limma**, linear modelling is carried out on the log-CPM values which are assumed to be normally distributed and the mean-variance relationship is accommodated using precision weights calculated by the `voom` function.

When operating on a DGEList-object, `voom` converts raw counts to log-CPM values by automatically extracting library sizes and normalisation factors from `x` itself. Additional normalisation to log-CPM values can be specified within `voom` using the `normalize.method` argument.

The mean-variance relationship of log-CPM values for this dataset is shown in Figure 4A. Typically, the "voom-plot" shows a decreasing trend between the means and variances resulting from a combination of technical variation in the sequencing experiment and biological variation amongst the replicate samples from different cell populations. Experiments with high biological variation usually result in flatter trends, where variance values plateau at high expression values. Experiments with low biological variation tend to result in sharp decreasing trends.

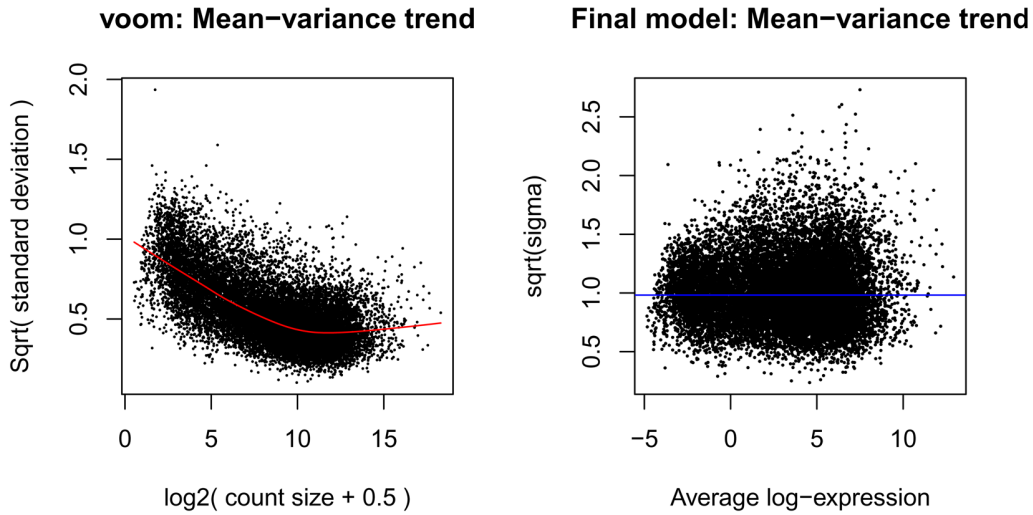


Figure 4. Means (x-axis) and variances (y-axis) of each gene are plotted to show the dependence between the two before `voom` is applied to the data (**A**) and how the trend is removed after `voom` precision weights are applied to the data (**B**). The plot on the left is created within the `voom` function which extracts residual variances from fitting linear models to log-CPM transformed data. Variances are then rescaled to quarter-root variances (or square-root of standard deviations) and plotted against the mean expression of each gene. The means are $\log_2$-transformed mean-counts with an offset of 2. The plot on the right is created using `plotSA` which plots $\log_2$ residual standard deviations against mean log-CPM values. The average $\log_2$ residual standard deviation is marked by a horizontal blue line. In both plots, each black dot represents a gene and a red curve is fitted to these points.

Moreover, the voom-plot provides a visual check on the level of filtering performed upstream. If filtering of lowly-expressed genes is insufficient, a drop in variance levels can be observed at the low end of the expression scale due to very small counts. If this is observed, one should return to the earlier filtering step and increase the expression threshold applied to the dataset.

Where sample-level variation is evident from earlier inspections of the MDS plot, the `voomWithQualityWeights` function can be used to simultaneously incorporate sample-level weights together with the abundance dependent weights estimated by `voom`[14]. For an example of this approach, see Liu *et al.* (2016)[15].

```
v <- voom(x, design, plot=TRUE)
v
```

```
## An object of class "EList"
## $genes
##    ENTREZID SYMBOL TXCHROM
## 1    497097   Xkr4    chr1
## 5     20671  Sox17    chr1
## 6     27395 Mrpl15    chr1
## 7     18777 Lypla1    chr1
## 9     21399  Tcea1    chr1
## 16619 more rows ...
##
## $targets
##                              files group lib.size norm.factors lane
## 10_6_5_11 GSM1545535_10_6_5_11.txt    LP 29387429        0.894 L004
## 9_6_5_11   GSM1545536_9_6_5_11.txt    ML 36212498        1.025 L004
## purep53     GSM1545538_purep53.txt Basal 59771061        1.046 L004
## JMS8-2       GSM1545539_JMS8-2.txt Basal 53711278        1.046 L006
```

```
## JMS8-3       GSM1545540_JMS8-3.txt    ML 77015912        1.016 L006
## JMS8-4       GSM1545541_JMS8-4.txt    LP 55769890        0.922 L006
## JMS8-5       GSM1545542_JMS8-5.txt Basal 54863512        0.996 L006
## JMS9-P7c   GSM1545544_JMS9-P7c.txt    ML 23139691        1.086 L008
## JMS9-P8c   GSM1545545_JMS9-P8c.txt    LP 19634459        0.984 L008
##
## $E
##         Samples
## Tags     10_6_5_11 9_6_5_11 purep53 JMS8-2 JMS8-3 JMS8-4 JMS8-5 JMS9-P7c JMS9-P8c
##   497097     -4.29    -3.86   2.519  3.293  -4.46  -3.99  3.287   -3.210    -5.30
##   20671      -4.29    -4.59   0.356 -0.407  -1.20  -1.94  0.844   -0.323    -1.21
##   27395       3.88     4.41   4.517  4.562   4.34   3.79  3.899    4.340     4.12
##   18777       4.71     5.57   5.396  5.162   5.65   5.08  5.060    5.751     5.14
##   21399       4.79     4.75   5.370  5.122   4.87   4.94  5.138    5.031     4.98
## 16619 more rows ...
##
## $weights
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,]  1.08  1.33  19.8 20.27  1.99  1.40 20.49  1.11  1.08
## [2,]  1.17  1.46   4.8  8.66  3.61  2.63  8.76  3.21  2.54
## [3,] 20.22 25.57  30.4 28.53 31.35 25.74 28.72 21.20 16.66
## [4,] 26.95 32.51  33.6 33.23 34.23 32.35 33.33 30.35 24.26
## [5,] 26.61 28.50  33.6 33.21 33.57 32.00 33.31 25.17 23.57
## 16619 more rows ...
##
## $design
##   Basal LP ML laneL006 laneL008
## 1     0  1  0        0        0
## 2     0  0  1        0        0
## 3     1  0  0        0        0
## 4     1  0  0        1        0
## 5     0  0  1        1        0
## 6     0  1  0        1        0
## 7     1  0  0        1        0
## 8     0  0  1        0        1
## 9     0  1  0        0        1
## attr(,"assign")
## [1] 1 1 1 2 2
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
##
## attr(,"contrasts")$lane
## [1] "contr.treatment"
```

Note that the other data frames stored within the DGEList-object that contain gene- and sample-level information, are retained in the EList-object `v` created by `voom`. The `v$genes` data frame is equivalent to `x$genes`, `v$targets` is equivalent to `x$samples`, and the expression values stored in `v$E` is analogous to `x$counts`, albeit on a transformed scale. In addition to this, the `voom` EList-object has a matrix of precision weights `v$weights` and stores the design matrix in `v$design`.

Fitting linear models for comparisons of interest

Linear modelling in **limma** is carried out using the `lmFit` and `contrasts.fit` functions originally written for application to microarrays. The functions can be used for both microarray and RNA-seq data and fit a separate model to the expression values for each gene. Next, empirical Bayes moderation is carried out by borrowing information across all genes to obtain more precise estimates of gene-wise variability[16]. The model's residual variances are plotted against average expression values in Figure 4B. It can be seen from this plot that the variance is no longer dependent on the mean expression level.

```
vfit <- lmFit(v, design)
vfit <- contrasts.fit(vfit, contrasts=contr.matrix)
efit <- eBayes(vfit)
plotSA(efit)
```

Examining the number of DE genes

For a quick look at differential expression levels, the number of significantly up- and down-regulated genes can be summarised in a table. Significance is defined using an adjusted *p*-value cutoff that is set at 5% by default. For the comparison between expression levels in basal and LP, 4,648 genes are found to be down-regulated in basal relative to LP and 4,863 genes are up-regulated in basal relative to LP – a total of 9,511 DE genes. A total of 9,598 DE genes are found between basal and ML (4,927 down- and 4,671 up-regulated genes), and a total of 5,652 DE genes are found between LP and ML (3,135 down- and 2,517 up-regulated). The larger numbers of DE genes observed for comparisons involving the basal population are consistent with our observations from the MDS plots.

```
summary(decideTests(efit))
```

```
##        BasalvsLP BasalvsML LPvsML
## Down        4648      4927   3135
## NotSig      7113      7026  10972
## Up          4863      4671   2517
```

Some studies require more than an adjusted *p*-value cutoff. For a stricter definition on significance, one may require log-fold-changes (log-FCs) to be above a minimum value. The *treat* method[17] can be used to calculate *p*-values from empirical Bayes moderated *t*-statistics with a minimum log-FC requirement. The number of differentially expressed genes are reduced to a total of 3,648 DE genes for basal versus LP, 3,834 DE genes for basal versus ML, and 414 DE genes for LP versus ML when testing requires genes to have a log-FC that is significantly greater than 1 (equivalent to a 2-fold difference between cell types on the original scale).

```
tfit <- treat(vfit, lfc=1)
dt <- decideTests(tfit)
summary(dt)
```

```
##        BasalvsLP BasalvsML LPvsML
## Down        1632      1777    224
## NotSig     12976     12790  16210
## Up          2016      2057    190
```

Genes that are DE in multiple comparisons can be extracted using the results from `decideTests`, where 0s represent genes that are not DE, 1s represent genes that are up-regulated, and -1s represent genes that are down-regulated. A total of 2,784 genes are DE in both basal versus LP and basal versus ML (Figure 5), twenty of which are listed below. The `write.fit` function can be used to extract and write results for all three comparisons to a single output file.

```
de.common <- which(dt[,1]!=0 & dt[,2]!=0)
length(de.common)
```

```
## [1] 2784
```

```
head(tfit$genes$SYMBOL[de.common], n=20)
```

```
##  [1] "Xkr4"         "Rgs20"         "Cpa6"           "A830018L16Rik" "Sulf1"
##  [6] "Eya1"         "Msc"           "Sbspon"         "Pi15"          "Crispld1"
## [11] "Kcnq5"        "Rims1"         "Khdrbs2"        "Ptpn18"        "Prss39"
## [16] "Arhgef4"      "Cnga3"         "2010300C02Rik"  "Aff3"          "Npas2"
```

```
vennDiagram(dt[,1:2], circle.col=c("turquoise", "salmon"))
write.fit(tfit, dt, file="results.txt")
```

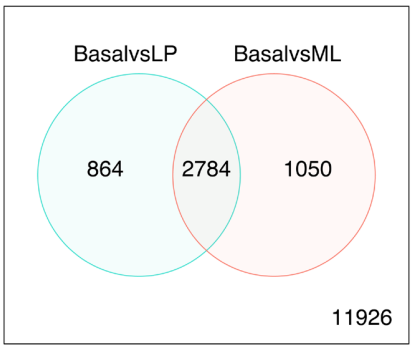


Figure 5. Venn diagram showing the number of genes DE in the comparison between basal versus LP only (left), basal versus ML only (right), and the number of genes that are DE in both comparisons (center). The number of genes that are not DE in either comparison are marked in the bottom-right.

Examining individual DE genes from top to bottom

The top DE genes can be listed using `topTreat` for results using `treat` (or `topTable` for results using `eBayes`). By default `topTreat` arranges genes from smallest to largest adjusted *p*-value with associated gene information, log-FC, average log-CPM, moderated *t*-statistic, raw and adjusted *p*-value for each gene. The number of top genes displayed can be specified, where `n=Inf` includes all genes. Genes *Cldn7* and *Rasef* are amongst the top DE genes for both basal versus LP and basal versus ML.

```
basal.vs.lp <- topTreat(tfit, coef=1, n=Inf)
basal.vs.ml <- topTreat(tfit, coef=2, n=Inf)
head(basal.vs.lp)
```

```
##        ENTREZID SYMBOL TXCHROM logFC AveExpr     t  P.Value adj.P.Val
## 12759     12759    Clu   chr14 -5.46    8.86 -33.6 1.72e-10  1.71e-06
## 53624     53624  Cldn7   chr11 -5.53    6.30 -32.0 2.58e-10  1.71e-06
## 242505   242505  Rasef    chr4 -5.94    5.12 -31.3 3.08e-10  1.71e-06
## 67451     67451   Pkp2   chr16 -5.74    4.42 -29.9 4.58e-10  1.74e-06
## 228543   228543   Rhov    chr2 -6.26    5.49 -29.1 5.78e-10  1.74e-06
## 70350     70350  Basp1   chr15 -6.08    5.25 -28.3 7.27e-10  1.74e-06
```

```
head(basal.vs.ml)
```

```
##        ENTREZID  SYMBOL TXCHROM logFC AveExpr     t  P.Value adj.P.Val
## 242505   242505   Rasef    chr4 -6.53    5.12 -35.1 1.23e-10  1.24e-06
## 53624     53624   Cldn7   chr11 -5.50    6.30 -31.7 2.77e-10  1.24e-06
## 12521     12521    Cd82    chr2 -4.69    7.07 -31.4 2.91e-10  1.24e-06
## 20661     20661   Sort1    chr3 -4.93    6.70 -30.7 3.56e-10  1.24e-06
## 71740     71740 Nectin4    chr1 -5.58    5.16 -30.6 3.72e-10  1.24e-06
## 12759     12759     Clu   chr14 -4.69    8.86 -28.0 7.69e-10  1.48e-06
```

Useful graphical representations of differential expression results

To summarise results for all genes visually, mean-difference plots, which display log-FCs from the linear model fit against the average log-CPM values can be generated using the `plotMD` function, with the differentially expressed genes highlighted.

```
plotMD(tfit, column=1, status=dt[,1], main=colnames(tfit)[1], xlim=c(-8,13))
```

Glimma extends this functionality by providing an interactive mean-difference plot via the `glMDPlot` function. The output of this function is an html page, with summarised results in the left panel (similar to what is output by `plotMD`), and the log-CPM values from individual samples for a selected gene in the right panel, with a table of results below the plots (Figure 6). This interactive display allows the user to search for particular genes based on the annotation provided (e.g. Gene symbol identifier), which is not possible in a static `R` plot.

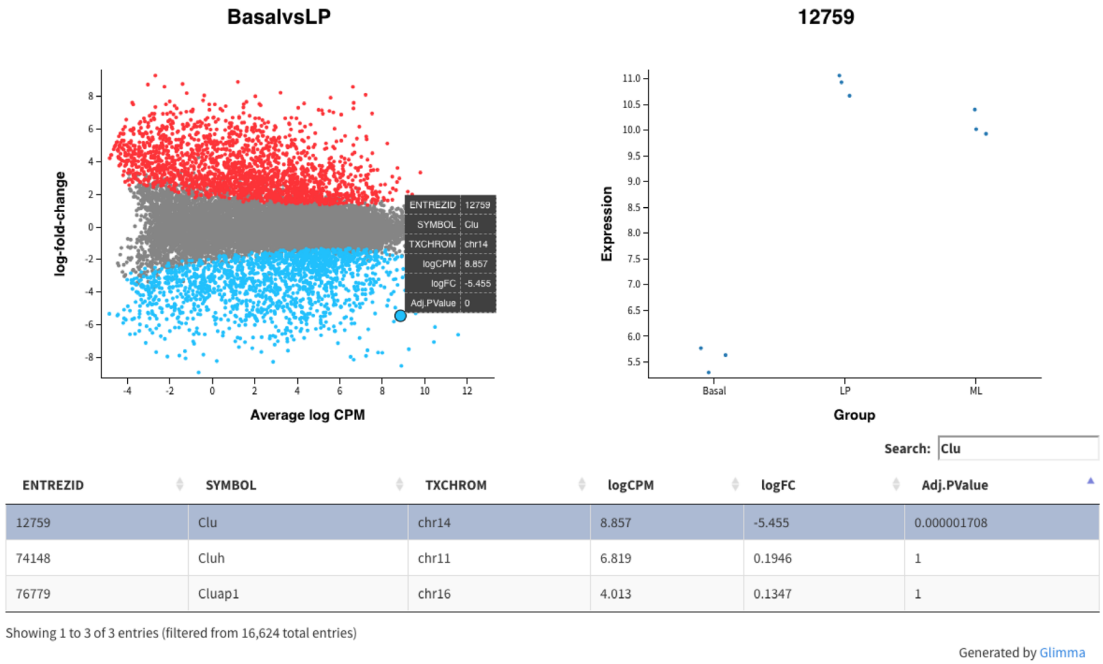


Figure 6. Interactive mean-difference plot generated using Glimma. Summary data (log-FCs versus log-CPM values) are shown in the left panel which is linked to the individual values per sample for a selected gene in the right panel. A table of results is also displayed below these figures, along with a search bar to allow users to look up a particular gene using the annotation information available, e.g. the Gene symbol identifier *Clu*.

```
glMDPlot(tfit, coef=1, status=dt, main=colnames(tfit)[1],
         side.main="ENTREZID", counts=lcpm, groups=group, launch=FALSE)
```

The mean-difference plot generated by the command above is available online (see http://bioinf.wehi.edu.au/folders/limmaWorkflow/glimma-plots/MD-Plot.html). The interactivity provided by the **Glimma** package allows additional information to be presented in a single graphical window. **Glimma** is implemented in `R` and Javascript, with the `R` code generating the data which is converted into graphics using the Javascript library D3 (`https://d3js.org`), with the Bootstrap library handling layouts and Datatables generating the interactive searchable tables. This allows plots to be viewed in any modern browser, which is convenient for including them as linked files from an Rmarkdown report of the analysis.

Plots shown previously include either all of the genes that are expressed in any one condition (such as the Venn diagram of common DE genes or mean-difference plot) or look at genes individually (log-CPM values shown in right panel of the interactive mean-difference plot). Heatmaps allow users to look at the expression of a subset of genes. This can give useful insight into the expression of individual groups and samples without losing perspective of the overall study when focusing on individual genes, or losing resolution when examining patterns averaged over thousands of genes at the same time.

A heatmap is created for the top 100 DE genes (as ranked by adjusted *p*-value) from the basal versus LP contrast using the `heatmap.2` function from the **gplots** package (Figure 7). The heatmap correctly clusters samples by cell type and reorders the genes into blocks with similar expression patterns. From the heatmap, we observe that the expression of ML and LP samples are very similar for the top 100 DE genes between basal and LP.

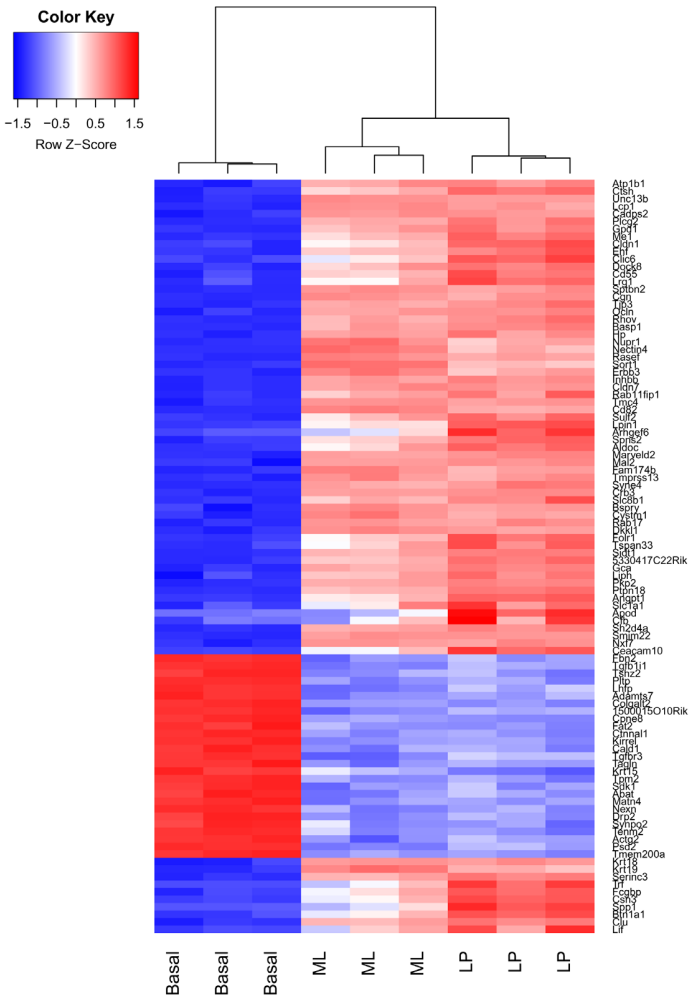


Figure 7. Heatmap of log-CPM values for top 100 genes DE in basal versus LP. Expression across each gene (or row) have been scaled so that mean expression is zero and standard deviation is one. Samples with relatively high expression of a given gene are marked in red and samples with relatively low expression are marked in blue. Lighter shades and white represent genes with intermediate expression levels. Samples and genes have been reordered by the method of hierarchical clustering. A dendrogram is shown for the sample clustering.

```
library(gplots)
basal.vs.lp.topgenes <- basal.vs.lp$ENTREZID[1:100]
i <- which(v$genes$ENTREZID %in% basal.vs.lp.topgenes)
mycol <- colorpanel(1000,"blue","white","red")

heatmap.2(lcpm[i,], scale="row",
   labRow=v$genes$SYMBOL[i], labCol=group,
   col=mycol, trace="none", density.info="none",
   margin=c(8,6), lhei=c(2,10), dendrogram="column")
```

Gene set testing with camera

We finish off this analysis with some gene set testing by applying the *camera* method[18] to the *c2* gene signatures from the Broad Institute's MSigDB c2 collection[19] that have been adapted for mouse and are available as Rdata objects from http://bioinf.wehi.edu.au/software/MSigDB/. Other useful gene sets derived from MSigDB for both human and mouse, such as the hallmark gene sets, are also available from this site. C2 gene sets have been curated

from online databases, publications and domain experts, and hallmark gene sets are selected from MSigDB to have well-defined biological states or processes.

```
load(url("http://bioinf.wehi.edu.au/software/MSigDB/mouse_c2_v5p1.rdata"))
idx <- ids2indices(Mm.c2,id=rownames(v))
cam.BasalvsLP <- camera(v,idx,design,contrast=contr.matrix[,1])
head(cam.BasalvsLP,5)
```

```
##                                             NGenes Direction    PValue       FDR
## LIM_MAMMARY_STEM_CELL_UP                       791        Up 1.77e-18 8.36e-15
## LIM_MAMMARY_STEM_CELL_DN                       683      Down 4.03e-14 8.69e-11
## ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER    170        Up 5.52e-14 8.69e-11
## LIM_MAMMARY_LUMINAL_PROGENITOR_UP               94      Down 2.74e-13 3.23e-10
## SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP         190        Up 5.16e-13 4.87e-10
```

```
cam.BasalvsML <- camera(v,idx,design,contrast=contr.matrix[,2])
head(cam.BasalvsML,5)
```

```
##                                      NGenes Direction    PValue       FDR
## LIM_MAMMARY_STEM_CELL_UP                791        Up 1.68e-22 7.92e-19
## LIM_MAMMARY_STEM_CELL_DN                683      Down 7.79e-18 1.84e-14
## LIM_MAMMARY_LUMINAL_MATURE_DN           172        Up 9.74e-16 1.53e-12
## LIM_MAMMARY_LUMINAL_MATURE_UP           204      Down 1.15e-12 1.36e-09
## NAKAYAMA_SOFT_TISSUE_TUMORS_PCA2_UP     137        Up 2.24e-12 1.88e-09
```

```
cam.LPvsML <- camera(v,idx,design,contrast=contr.matrix[,3])
head(cam.LPvsML,5)
```

```
##                                          NGenes Direction    PValue       FDR
## LIM_MAMMARY_LUMINAL_MATURE_DN               172        UP 6.73e-14 2.35e-10
## LIM_MAMMARY_LUMINAL_MATURE_UP               204      Down 9.97e-14 2.35e-10
## LIM_MAMMARY_LUMINAL_PROGENITOR_UP            94        Up 1.32e-11 2.08e-08
## REACTOME_RESPIRATORY_ELECTRON_TRANSPORT      94      Down 7.01e-09 8.28e-06
## REACTOME_RNA_POL_I_PROMOTER_OPENING          46      Down 2.04e-08 1.93e-05
```

The `camera` function performs a competitive test to assess whether the genes in a given set are highly ranked in terms of differential expression relative to genes that are not in the set. It uses **limma**'s linear model framework, taking both the design matrix and contrast matrix (if present) and accommodates the observational-level weights from *voom* in the testing procedure. After adjusting the variance of the resulting gene set test statistic by a variance inflation factor that depends on the gene-wise correlation (which is set to 0.01 by default, but can be estimated from the data) and the size of the set, a *p*-value is returned and adjusted for multiple testing.

This experiment is the RNA-seq equivalent of the dataset generated by Lim *et al.* (2010)[20], who used Illumina microarrays to profile the same sorted cell populations, so it is reassuring to see the gene signatures from this earlier publication coming up at the top of the list for each contrast. We make a barcodeplot of the Lim *et al.* (2010) Mature Luminal gene sets (Up and Down) in the LP versus ML contrast. Note that these sets go in the opposite direction in our dataset due to our parameterization which compares LP against ML rather than the other way around (if the contrast were reversed, the directions would be consistent).

```
barcodeplot(efit$t[,3], index=idx$LIM_MAMMARY_LUMINAL_MATURE_UP,
            index2=idx$LIM_MAMMARY_LUMINAL_MATURE_DN, main="LPvsML")
```

Other gene set tests are available in **limma**, such as the self-contained tests by *mroast*[21]. Whilst *camera* is ideal for testing a large database of gene sets and observing which of them rank highly relative to others (as shown above), self-contained tests are better for focused testing of one or a few specifically chosen sets to see if they are DE in their own right. In other words, *camera* is more appropriate when "fishing" for gene sets of interest, whereas *mroast* tests sets that are already of interest for significance.

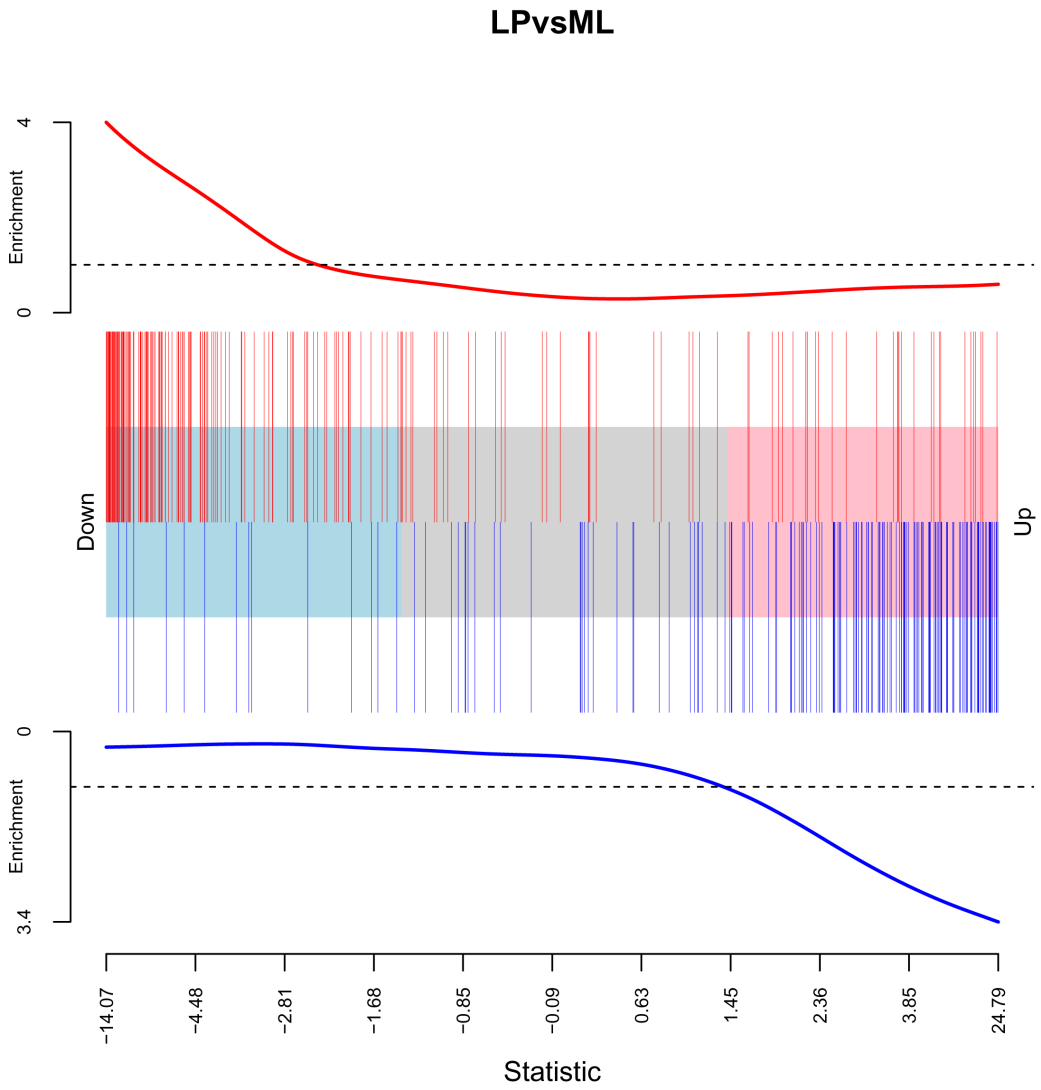


Figure 8. Barcode plot of LIM_MAMMARY_LUMINAL_MATURE_UP (red bars, top of plot) and LIM_MAMMARY_LUMINAL_MATURE_DN (blue bars, bottom of plot) gene sets in the LP versus ML contrast. For each set, an enrichment line that shows the relative enrichment of the vertical bars in each part of the plot is displayed. The experiment of Lim *et al.* (2010) is very similar to the current one, with the same sorting strategy used to obtain the different cell populations, except that microarrays were used instead of RNA-seq to profile gene expression. Note that the inverse correlation (the up gene set is down and the down gene set is up) is a result of the way the contrast has been set up (LP versus ML) – if reversed, the directionality would agree.

Software availability

This RNA-seq workflow makes use of various packages available from version 3.8 of the Bioconductor project, running on R[22] version 3.5.1. Besides the software highlighted in this article (**limma**, **Glimma** and **edgeR**) it requires a number of other packages, including **gplots**[23] and **RColorBrewer** and the gene annotation package **Mus.musculus**. This document was compiled using **knitr**[24–26]. Version numbers for all packages used are shown below. The **RNAseq123** Bioconductor workflow package available from `https://bioconductor.org/packages/RNAseq123` contains both an English and Chinese (Mandarin) vignette of this article along with code to perform the complete analysis. Installation of this package manages all of the above-mentioned dependencies and is a useful resource for delivering hands-on training on RNA-seq data analysis.

```
sessionInfo()

## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_AU.UTF-8/en_AU.UTF-8/en_AU.UTF-8/C/en_AU.UTF-8/en_AU.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics grDevices utils     datasets methods
## [9] base
##
## other attached packages:
##  [1] gplots_3.0.1                        RColorBrewer_1.1-2
##  [3] Mus.musculus_1.3.1                  TxDb.Mmusculus.UCSC.mm10.knownGene_3.4.4
##  [5] org.Mm.eg.db_3.7.0                  GO.db_3.7.0
##  [7] OrganismDbi_1.24.0                  GenomicFeatures_1.34.1
##  [9] GenomicRanges_1.34.0                GenomeInfoDb_1.18.1
## [11] AnnotationDbi_1.44.0                IRanges_2.16.0
## [13] S4Vectors_0.20.1                    Biobase_2.42.0
## [15] BiocGenerics_0.28.0                 edgeR_3.24.0
## [17] Glimma_1.10.0                       limma_3.38.3
## [19] knitr_1.20                          BiocStyle_2.10.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.3.1                  bit64_0.9-7                 jsonlite_1.5
##  [4] R.utils_2.7.0               gtools_3.8.1                assertthat_0.2.0
##  [7] BiocManager_1.30.4          highr_0.7                   RBGL_1.58.1
## [10] blob_1.1.1                  GenomeInfoDbData_1.2.0      Rsamtools_1.34.0
## [13] yaml_2.2.0                  progress_1.2.0              RSQLite_2.1.1
## [16] backports_1.1.2             lattice_0.20-38             digest_0.6.18
## [19] XVector_0.22.0              htmltools_0.3.6             Matrix_1.2-15
## [22] R.oo_1.22.0                 XML_3.98-1.16               pkgconfig_2.0.2
## [25] biomaRt_2.38.0              zlibbioc_1.28.0             gdata_2.18.0
## [28] BiocParallel_1.16.2         SummarizedExperiment_1.12.0 magrittr_1.5
## [31] crayon_1.3.4                memoise_1.1.0               evaluate_0.12
## [34] R.methodsS3_1.7.1           graph_1.60.0                tools_3.5.1
## [37] prettyunits_1.0.2           hms_0.4.2                   matrixStats_0.54.0
## [40] stringr_1.3.1               locfit_1.5-9.1              DelayedArray_0.8.0
## [43] Biostrings_2.50.1           compiler_3.5.1              caTools_1.17.1.1
## [46] rlang_0.3.0.1               grid_3.5.1                  RCurl_1.95-4.11
## [49] bitops_1.0-6                rmarkdown_1.10              DBI_1.0.0
## [52] R6_2.3.0                    GenomicAlignments_1.18.0    rtracklayer_1.42.1
## [55] bit_1.1-14                  rprojroot_1.3-2             KernSmooth_2.23-15
## [58] stringi_1.2.4               Rcpp_1.0.0
```

Author contributions

CWL, GKS and MER developed the workflow and wrote the article with input from MA (gene set testing), SS (Glimma interactive plotting), XD (Chinese translation of vignette and all code and figure updates required for Bioconductor 3.8) and LT (Chinese translation of vignette).

Grant information

This worked was funded by the National Health and Medical Research Council (NHMRC) (Fellowship GNT1058892 and Program GNT1054618 to GKS, Project GNT1050661 to MER and GKS and Fellowship GNT1104924 to

MER), Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Dr Julie Sheridan for generating this dataset and for advice on its analysis.

References

1. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1):139–40.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Ritchie ME, Phipson B, Wu D, *et al.*: ***limma* powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; **43**(7): e47.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–21.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Su S, Law CW, Ah-Cann C, *et al.*: **Glimma: interactive graphics for gene expression analysis.** *Bioinformatics.* 2017; **33**(13):2050–2052.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Sheridan JM, Ritchie ME, Best SA, *et al.*: **A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for *Asap1* and *Prox1*.** *BMC Cancer.* 2015; **15**(1): 221.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Liao Y, Smyth GK, Shi W: **The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.** *Nucleic Acids Res.* 2013; **41**(10):e108.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**(7): 923–30.
PubMed Abstract | Publisher Full Text
8. Bioconductor Core Team: **Mus.musculus: Annotation package for the Mus.musculus object.** R package version 1.3.1. 2016.
Publisher Full Text
9. Bioconductor Core Team: **Homo.sapiens: Annotation package for the Homo.sapiens object.** R package version 1.3.1. 2016.
Publisher Full Text
10. Durinck S, Moreau Y, Kasprzyk A, *et al.*: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics.* 2005; **21**(16):3439–40.
PubMed Abstract | Publisher Full Text
11. Durinck S, Spellman PT, Birney E, *et al.*: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc.* 2009; **4**(8): 1184–91.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol.* 2010; **11**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Law CW, Chen Y, Shi W, *et al.*: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol.* 2014; **15**(2): R29.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Liu R, Holik AZ, Su S, *et al.*: **Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses.** *Nucleic Acids Res.* 2015; **43**(15): e97.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Liu R, Chen K, Jansz N, *et al.*: **Transcriptional profiling of the epigenetic regulator Smchd1.** *Genom Data.* 2016; **7**: 144–7.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol.* 2004; **3**(1): Article3.
PubMed Abstract | Publisher Full Text
17. McCarthy DJ, Smyth GK: **Testing significance relative to a fold-change threshold is a TREAT.** *Bioinformatics.* 2009; **25**(6): 765–71.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation.** *Nucleic Acids Res.* 2012; **40**(17): e133.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–50.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Lim E, Wu D, Pal B, *et al.*: **Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways.** *Breast Cancer Res.* 2010; **12**(2): R21.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Wu D, Lim E, Vaillant F, *et al.*: **ROAST: rotation gene set tests for complex microarray experiments.** *Bioinformatics.* 2010; **26**(17): 2176–82.
PubMed Abstract | Publisher Full Text | Free Full Text
22. R Development Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, 2016.
Reference Source
23. Warnes GR, Bolker B, Bonebakker L, *et al.*: **gplots: Various R Programming Tools for Plotting Data.** R package version 3.0.1. 2016.
Reference Source
24. Xie Y: **knitr: A comprehensive tool for reproducible research in R.** In: V. Stodden, F. Leisch, and R. D. Peng, editors, *Implementing Reproducible Computational Research.* Chapman and Hall/CRC, 2014; ISBN 978-1466561595.
Reference Source
25. Xie Y: **Dynamic Documents with R and knitr.** Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015; ISBN 978-1498716963.
Reference Source
26. Xie Y: **knitr: A General-Purpose Package for Dynamic Report Generation in R.** R package version 1.12.3. 2016.
Reference Source

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

Referee Report 06 July 2016

https://doi.org/10.5256/f1000research.9688.r14437

✔ **Jovana Maksimovic**
Bioinformatics research group, Murdoch Childrens Research Institute, Murdoch Children's Research Institute, Parkville, Vic, Australia

This workflow outlines a step-by-step differential expression analysis of a publicly available RNA-seq dataset using several well-established Bioconductor packages such as limma and edgeR and the novel Glimma package. The workflow describes in detail the key steps that are generally performed as part of most standard, count-based RNAseq analyses: pre-processing, exploratory data analysis, differential expression testing and pathway analysis. Each of the steps is carefully explained and broken down into a series of logical substeps, which also include R code and example output. Furthermore, this workflow is a good example of how multiple Bioconductor packages can be applied in a real-world RNAseq analysis.

Overall, I think this is a very clearly written and well-explained workflow that not only highlights the utility of combining Bioconductor packages for RNAseq analysis but also demystifies the process for those who may be interested in learning what an RNAseq analysis involves. Although some familiarity with R and statistical concepts would be advantageous, they are not necessary to be able to follow the logic behind the steps and what each step is trying to achieve.

I have tested the R code under R 3.3.0 and Bioconductor 3.0 and it ran without errors.

Minor comments:

- As this workflow will be of particular use for those who are very new to RNAseq analysis, I think it should be explained why 1 CPM is chosen as the cutoff for lowly expressed genes. In addition, it may be useful to show the relationship between CPM and raw counts and how the selected cutoff relates to numbers of reads.
- In the gene set testing section, it is worth mentioning that the Hallmark gene sets (also available from the Broad MSigDB) may be a good starting point for pathway analysis as they “summarise and represent specific well-defined, states or processes”.
- Also, for those who are new to RNAseq analysis, it may be worth mentioning that there are other gene set testing methods available in limma e.g. roast etc. and a very brief explanation of when the different methods are appropriate.
- General comment regarding Glimma plots: it would be handy to see the gene symbol (if available) as well as EntrezID in the title of the scatter plot next to the MD plot.

Competing Interests: Matthew E. Ritchie and I are both organisers of the Bioconductor Asia-Pacific Developer meeting, which is to be held in Brisbane in November 2016.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 30 Nov 2016

Charity Law, The Walter and Eliza Hall Institute of Medical Research, Australia

Dear Jovana, thank-you for the helpful suggestions. We have added extra text to explain why a CPM value of 1 was chosen and how this value relates to raw counts. We have also given more detail on c2 and Hallmark gene sets, as well as mentioning roast and when one might chose to use roast over camera.

Competing Interests: None

Referee Report 06 July 2016

https://doi.org/10.5256/f1000research.9688.r14439

Maria A. Doyle
Research Computing Facility, Peter MacCallum Cancer Centre, Melbourne, Vic, Australia

Editorial Note from F1000Research - 7th July 2016:

the status of this review was changed after publication from "Approved with Reservations" to "Approved". The referee's reservation was caused by using the current release (Bioconductor 3.3) as opposed to the development release (Bioconductor 3.4) as used in the article.

The following text has been removed:

My only reservation to approving the article is that I encountered these two errors when running the Glimma code in the article. I used R 3.3.0 and the most recent version of Glimma from Bioconductor (Glimma 1.0.0) but the sessionInfo() in the article shows Glimma 1.1.1 was used which may account for the errors:

1) With this line of code no plot was produced and it gave the error below:

```
glMDSPlot(lcpm, labels=paste(group, lane, sep="_"),
+         groups=x$samples[,c(2,5)], launch=TRUE)
Error in glScatter.default(points, xval = "dim1", yval = "dim2", xlab = "Dimension 1",  :
  group does not correspond to a column
```

2) With this line of code the table (shown in Fig 6.) didn't appear, just the 2 plots:

```
glMDPlot(tfit, coef=1, status=dt[,1], main=colnames(tfit)[1],
   counts=x$counts, samples=colnames(x), anno=x$genes, groups=group,
   id.column="ENTREZID", display.columns=c("SYMBOL", "ENTREZID"),
   search.by="SYMBOL", launch=FALSE)
```

This article is a very nice description (nicely written, nicely explained) of an RNA-seq workflow, starting from gene counts and proceeding through a standard differential expression analysis, using 3

complementary Bioconductor packages edgeR, limma & Glimma. It explains clearly how to perform the steps of a differential expression analysis and also how to generate useful plots for visualising the data e.g. MDSplot, MDplot, barcodeplot, heatmap etc.

Two of the packages, edgeR and limma, are well established while Glimma is the new kid on the block. It's a tool along the lines of the web-based Degust (http://www.vicbioinformatics.com/degust/), in that it enables interactive exploration of RNA-seq data. It would seem to be a very useful addition to R RNA-seq workflows as the interactive html plots it can generate (MDS, MD etc) make exploring the results easier (especially by non-R savvy collaborators) and it could help save an analyst's time and effort through not having to reproduce static plots e.g. to highlight different genes. With the Glimma MDplot I like that you can search for a gene and see a plot of the log-cpm counts for that gene in the samples.

It was great to be able to try out the workflow really easily by downloading the data file linked to in the paper (from GEO), with no processing required other than unzipping the files, and the code in the article all worked, giving the same results shown in the paper with the exception of two errors described below.

Minor comments:

The article says the workflow is available from Bioconductor here http://www.bioconductor.org/help/workflows/ but it doesn't seem to be there, however it is available at the other location mentioned http://bioinf.wehi.edu.au/folders/limmaWorkflow/

The section title "Removing genes that are not expressed" suggests only genes with no expression are removed, however the paragraph then explains that genes lowly expressed are also removed so for greater clarity maybe that could be changed to something like "Removing genes that are not sufficiently expressed".

In the section "Organising gene annotations" I think it's good the authors point out to check for duplicated genes, however in this case the duplication appears to be due to extracting the TXCHROM column (as some genes are reported as being present on more than one chromosome) but the TXCHROM information is not used in this workflow and if the TXCHROM column is omitted then there are no duplicates so it might be worth mentioning that.

With this bit "Differential expression analyses look at gene expression differences between conditions, rather than comparing expression across multiple genes or drawing conclusions on absolute levels of expression. In other words, gene lengths remain constant for comparisons of interest and any observed differences are a result of changes in condition rather than changes in gene length." Wouldn't one caveat to this be if there was a significant change in the length of the isoform(s) expressed from a gene (e.g. from expression of a long isoform to a short isoform) as then the assumption of no change in gene length would no longer be valid.

In the online version Fig.7 is in the middle of a code block could it be moved below.

Figs 5-8 have bold headings but Figs 1-4 don't is that an error.

Could perhaps modify title to reflect the order in which the tools are used - edger, limma & Glimma.

It would be nice for consistency if the colours in the Glimma interactive MDS plots matched the colours used in the static MDS plots in Fig 3.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Response 06 Jul 2016

Maria A. Doyle, Peter MacCallum Cancer Centre, Australia

I've just requested that the status of this article be changed to Approved as the code works fine with the development release of Bioconductor 3.4 but could it be made clearer in the article that 3.4 must be used for the code to work as shown.

Competing Interests: No competing interests were disclosed.

Author Response 30 Nov 2016

Charity Law, The Walter and Eliza Hall Institute of Medical Research, Australia

Dear Maria, we have made changes according to your suggestions, including a change in the section title to "Removing genes that are lowly expressed", and we have added comments to point out that our analysis assumes no differential isoform usage. We have also simplified our example in the "Organising gene annotations" section.

Thanks for pointing out the inconsistencies in the figure captions. They have now been addressed.

Indeed, it would be nice to make the colours consistent between the interactive and static MDS plots. We like the colours used in the static plot and can have the interactive version of the plot updated once color specification becomes available to the MDS plotting function in Glimma.

Competing Interests: None

Referee Report 23 June 2016

https://doi.org/10.5256/f1000research.9688.r14436

James W. MacDonald
Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, USA

In this manuscript the authors perform a step-by-step analysis of a public RNA-Seq data set, using the Bioconductor packages edgeR, limma, and Glimma. The analysis proceeds through four stages (data preparation, preprocessing, univariate analyses, and gene set testing), with each stage broken up into several discrete steps. Each step is clearly explained and is accompanied by R code and output, so the reader can follow along if desired.

The target audience for this paper appears to be someone with passing familiarity with both R and statistics, without requiring expertise in either. Most of the code should be easy to understand, and any

complex sections are clearly explained in the text. Similarly, the authors outline the analytic choices they make (e.g., filtering out unexpressed genes, specifying model coefficients, etc) at a more accessible level.

While both edgeR and limma are well known, popular Bioconductor packages, Glimma is a new package that was released in April 2016. This package uses the d3.js JavaScript library to generate interactive HTML documents that can be viewed locally, rather than needing to be accessed from a server running R (as, say a shiny app requires). This is an exciting development, and is unfortunately not as compelling as it could be, if the Glimma plots were part of the HTML version of the manuscript rather than provided as links.

This is a well written paper, and is a useful contribution to the literature; while each package has either an extensive user's guide or vignette, by necessity these documents pertain only to the package at hand. Most RNA-Seq analyses require a combination of multiple packages to complete, and this paper provides a clear example.

Major comments

None.

Minor comments

In the section 'Organising gene annotations', the code used to subset the one-to-many mappings is needlessly complex. Simply doing something like

genes <- genes[!duplicated(genes[,1]),]

will accomplish the same thing, in a more straightforward way.

In the section 'Removing heteroscedasticity from count data', the authors state:

"When operating on a DGEList-object, voom converts raw counts to log-CPM values by automatically extracting library sizes and normalisation factors stored in the object. For a matrix of counts, the method of normalisation can be specified within voom using the normalize.method (by default no normalisation is performed)."

This is confusing; the authors already showed in an earlier section ('Normalising gene expression distributions') that converting to log-CPM using TMM normalization factors will do a shift-normalization (which is what voom will do in this instance). The normalize.method argument to voom specifies **additional** normalization methods that can be applied to the matrix of TMM normalized log-CPM values.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 30 Nov 2016

Charity Law, The Walter and Eliza Hall Institute of Medical Research, Australia

Dear James, thank-you for your comments. We have considered your suggestions and made changes to simplify the section on 'Organising gene annotations' and to clarify that voom offers additional normalisation ontop of the TMM-normsalisation that is carried out.

Competing Interests: None

Discuss this Article

Version 2

Reader Comment 15 Feb 2017

Marc Teunis, University of Applied Sciences, Utrecht, The Netherlands

Dear authors,

I am testing the workflow on a CentOS RStudio Server (see session info below). I just wanted to point out a minor glitch in the code for downloading the files for this workflow.

I have installed the workflow from BIOCONDUCTOR using:

source("http://bioconductor.org/workflows.R")
workflowInstall("RNAseq123")

The current code chunck for downloading the file should be changed to

utils::download.file(url, destfile="GSE63310_RAW.tar", mode="wb", method = "wget")

Not setting method to "wget" results in an incomplete download of the RAW.tar archive.

Hoping to have aided future users with this comment.

Sincerely,
Marc

output sessionInfo() (partly)

R version 3.3.2 (2016-10-31)
Platform: x86_64-redhat-linux-gnu (64-bit)
Running under: CentOS Linux 7 (Core)

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C

[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

...

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research

D.8 Proteomics

Further explanations for Chapter 8.

D.8.1 Mass spectrometry methods

- Dupree et al. (2020) — Walks through the bottom-up workflow from sample preparation and digestion to LC-MS/MS, acquisition and quantification, and discusses its limitations. A readable overview of the section's experimental side. *(review; CC BY 4.0)*

  A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Dupree et al. (2020). *Proteomes*. doi:10.3390/proteomes8030014

  > Proteomics is the field of study that includes the analysis of proteins, from either a basic science prospective or a clinical one. Proteins can be investigated for their abundance, variety of proteoforms due to post-translational modifications (PTMs), and their stable or transient protein-protein interactions. This can be especially beneficial in the clinical setting when studying proteins involved in different diseases and conditions. Here, we aim to describe a bottom-up proteomics workflow from sample preparation to data analysis, including all of its benefits and pitfalls. We also describe potential improvements in this type of proteomics workflow for the future.

- Wang et al. (2024) — Tutorial on multiplex isotope labelling for quantification (precursor-based labels such as SILAC, mass-defect labels and reporter-ion labels such as TMT), with their principles, advantages and limitations such as ratio distortion. Useful background for the quantification subsection. *(tutorial; CC BY 4.0)*

  A Tutorial Review of Labeling Methods in Mass Spectrometry-Based Quantitative Proteomics. Wang et al. (2024). *ACS Measurement Science Au*. doi:10.1021/acsmeasuresciau.4c00007

  > Recent advancements in mass spectrometry (MS) have revolutionized quantitative proteomics, with multiplex isotope labeling emerging as a key strategy for enhancing accuracy, precision, and throughput. This tutorial review offers a comprehensive overview of multiplex isotope labeling techniques, including precursor-based, mass defect-based, reporter ion-based, and hybrid labeling methods. It details their fundamental principles, advantages, and inherent limitations along with strategies to mitigate the limitation of ratio-distortion. This review will also cover the applications and latest progress in these labeling techniques across various domains, including cancer biomarker discovery, neuroproteomics, post-translational modification analysis, cross-linking MS, and single-cell proteomics. This Review aims to provide guidance for researchers on selecting appropriate methods for their specific goals while also highlighting the potential future directions in this rapidly evolving field.

- Jiang et al. (2024) — Long, comprehensive review covering instruments, fragmentation, acquisition modes, data analysis and quantification in bottom-up proteomics. Useful as a reference to dip into; licence allows linking only. *(review; CC BY-NC-ND 4.0 (link only))*

  Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry. Jiang et al. (2024). *ACS Measurement Science Au*. doi:10.1021/acsmeasuresciau.3c00068

  > Proteomics is the large scale study of protein structure and function from biological systems through protein identification and quantification. "Shotgun proteomics" or "bottom-up proteomics" is the prevailing strategy, in which proteins are hydrolyzed into peptides that are analyzed by mass spectrometry. Proteomics studies can be applied to diverse studies ranging from simple protein identification to studies of proteoforms, protein-protein interactions, protein structural alterations, absolute and relative protein quantification, post-translational modifications, and protein stability. To enable this range of different experiments, there are diverse strategies for proteome analysis. The nuances of how proteomic workflows differ may be challenging to understand for new practitioners. Here, we provide a comprehensive overview of different proteomics methods. We cover from biochemistry basics and protein extraction to biological interpretation and orthogonal validation. We expect this Review will serve as a handbook for researchers who are new to the field of bottom-up proteomics.

D.8.2 Algorithms for peptide identification

- McHugh and Arthur (2008) — Short introductory review of peptide mass fingerprinting, database search, de novo sequencing and their scoring, written for computational readers. Older, but the basic algorithms it explains have not changed. *(review; CC BY 4.0)*

  Computational Methods for Protein Identification from Mass Spectrometry Data. McHugh & Arthur (2008). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.0040012

  > Protein identification using mass spectrometry is an indispensable computational tool in the life sciences. A dramatic increase in the use of proteomic strategies to understand the biology of living systems generates an ongoing need for more effective, efficient, and accurate computational methods for protein identification. A wide range of computational methods, each with various implementations, are available to complement different proteomic approaches. A solid knowledge of the range of algorithms available and, more critically, the accuracy and effectiveness of these techniques is essential to ensure as many of the proteins as possible, within any particular experiment, are correctly identified. Here, we undertake a systematic review of the currently available methods and algorithms for interpreting, managing, and analyzing biological data associated with protein identification. We summarize the advances in computational solutions as they have responded to corresponding advances in mass spectrometry hardware. The evolution of scoring algorithms and metrics for automated protein identification are also discussed with a focus on the relative performance of different techniques. We also consider the relative advantages and limitations of different techniques in particular biological contexts. Finally, we present our perspective on future developments in the area of computational protein identification by considering the most recent literature on new and promising approaches to the problem as well as identifying areas yet to be explored and the potential application of methods from other areas of computational biology.

- Angelis et al. (2025) — Introduction to deep-learning models that predict fragment spectra, retention times and other peptide properties, and how they are used in rescoring and library generation. Matches the subsection on predicting spectra and retention times. *(tutorial; CC BY 4.0)*

  Peptide Property Prediction for Mass Spectrometry Using AI: An Introduction to State of the Art Models. Angelis et al. (2025). *PROTEOMICS*. doi:10.1002/pmic.202400398

  > This review explores state of the art machine learning and deep learning models for peptide property prediction in mass spectrometry-based proteomics, including, but not limited to, models for predicting digestibility, retention time, charge state distribution, collisional cross section, fragmentation ion intensities, and detectability. The combination of these models enables not only the in silico generation of spectral libraries but also finds many additional use cases in the design of targeted assays or data-driven rescoring. This review serves as both an introduction for newcomers and an update for experienced researchers aiming to develop accessible and reproducible models for peptide property predictions. Key limitations of the current models, including difficulties in handling diverse post-translational modifications and instrument variability, highlight the need for large-scale, harmonized datasets, and standardized evaluation metrics for benchmarking.

- Wen et al. (2020) — Overview of deep-learning applications in proteomics, including spectrum and retention-time prediction and de novo peptide sequencing. Gives a broader context for the machine-learning methods in this section. *(review; CC BY 4.0)*

  Deep Learning in Proteomics. Wen et al. (2020). *PROTEOMICS*. doi:10.1002/pmic.201900335

  > Proteomics, the study of all the proteins in biological systems, is becoming a data-rich science. Protein sequences and structures are comprehensively catalogued in online databases. With recent advancements in tandem mass spectrometry (MS) technology, protein expression and post-translational modifications (PTMs) can be studied in a variety of biological systems at the global scale. Sophisticated computational algorithms are needed to translate the vast amount of data into novel biological insights. Deep learning automatically extracts data representations at high levels of abstraction from data, and it thrives in data-rich scientific research domains. Here, a comprehensive overview of deep learning applications in proteomics, including retention time prediction, MS/MS spectrum prediction, de novo peptide sequencing, PTM prediction, major histocompatibility complex-peptide binding prediction, and protein structure prediction, is provided. Limitations and the future directions of deep learning in proteomics are also

discussed. This review will provide readers an overview of deep learning and how it can be used to analyze proteomics data.

D.8.3 Statistics and error control

- Noble and MacCoss (2012) — A chapter of the PLoS Computational Biology 'Translational Bioinformatics' collection that explains peptide identification, target–decoy FDR estimation, q-values, posterior error probabilities, protein inference and quantification. Written by researchers behind Percolator and Skyline, it is a clear statistical introduction to the whole section. *(book chapter; CC BY 4.0)*

  Computational and Statistical Analysis of Protein Mass Spectrometry Data. Noble & MacCoss (2012). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002296

  > High-throughput proteomics experiments involving tandem mass spectrometry produce large volumes of complex data that require sophisticated computational analyses. As such, the field offers many challenges for computational biologists. In this article, we briefly introduce some of the core computational and statistical problems in the field and then describe a variety of outstanding problems that readers of PLoS Computational Biology might be able to help solve.

- Wen et al. (2025) — Explains how entrapment experiments can test whether a pipeline's reported FDR is valid, and shows that some widely used tools do not control the FDR as claimed. Useful for understanding the assumptions behind target–decoy estimates. *(perspective; CC BY 4.0)*

  Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. Wen et al. (2025). *Nature Methods*. doi:10.1038/s41592-025-02719-x

  > A critical challenge in mass spectrometry proteomics is accurately assessing error control, especially given that software tools employ distinct methods for reporting errors. Many tools are closed-source and poorly documented, leading to inconsistent validation strategies. Here we identify three prevalent methods for validating false discovery rate (FDR) control: one invalid, one providing only a lower bound, and one valid but under-powered. The result is that the proteomics community has limited insight into actual FDR control effectiveness, especially for data-independent acquisition (DIA) analyses. We propose a theoretical framework for entrapment experiments, allowing us to rigorously characterize different approaches. Moreover, we introduce a more powerful evaluation method and apply it alongside existing techniques to assess existing tools. We first validate our analysis in the better-understood data-dependent acquisition setup, and then, we analyze DIA data, where we find that no DIA search tool consistently controls the FDR, with particularly poor performance on single-cell datasets.

- Hutchings et al. (2023) — Step-by-step Bioconductor workflow for TMT and label-free data, covering import, pre-processing, quality control, aggregation to proteins and differential abundance testing. A practical complement to the statistics of quantification. *(tutorial; CC BY 4.0)*

  A Bioconductor workflow for processing, evaluating, and interpreting expression proteomics data. Hutchings et al. (2023). *F1000Research*. doi:10.12688/f1000research.139116.1

  > Background: Expression proteomics involves the global evaluation of protein abundances within a system. In turn, differential expression analysis can be used to investigate changes in protein abundance upon perturbation to such a system. Methods: Here, we provide a workflow for the processing, analysis and interpretation of quantitative mass spectrometry-based expression proteomics data. This workflow utilizes open-source R software packages from the Bioconductor project and guides users end-to-end and step-by-step through every stage of the analyses. As a use-case we generated expression proteomics data from HEK293 cells with and without a treatment. Of note, the experiment included cellular proteins labelled using tandem mass tag (TMT) technology and secreted proteins quantified using label-free quantitation (LFQ). Results: The workflow explains the software infrastructure before focusing on data import, pre-processing and quality control. This is done individually for TMT and LFQ datasets. The application of statistical differential expression analysis is demonstrated, followed by interpretation via gene ontology enrichment analysis. Conclusions: A comprehensive workflow for the processing, analysis and interpretation of expression proteomics is presented. The workflow is a valuable resource for the

proteomics community and specifically beginners who are at least familiar with R who wish to understand and make data-driven decisions with regards to their analyses.

D.8.4 Proteomics beyond protein abundance

- Richards et al. (2021) — Reviews affinity purification–MS, proximity labelling, co-fractionation and cross-linking MS for mapping interaction networks, including scoring and quality control. Covers the interaction proteomics subsection. *(review; CC BY 4.0)*

  Mass spectrometry-based protein–protein interaction networks for the study of human diseases. Richards et al. (2021). *Molecular Systems Biology*. doi:10.15252/msb.20188792

  > A better understanding of the molecular mechanisms underlying disease is key for expediting the development of novel therapeutic interventions. Disease mechanisms are often mediated by interactions between proteins. Insights into the physical rewiring of protein-protein interactions in response to mutations, pathological conditions, or pathogen infection can advance our understanding of disease etiology, progression, and pathogenesis and can lead to the identification of potential druggable targets. Advances in quantitative mass spectrometry (MS)-based approaches have allowed unbiased mapping of these disease-mediated changes in protein-protein interactions on a global scale. Here, we review MS techniques that have been instrumental for the identification of protein-protein interactions at a system-level, and we discuss the challenges associated with these methodologies as well as novel MS advancements that aim to address these challenges. An overview of examples from diverse disease contexts illustrates the potential of MS-based protein-protein interaction mapping approaches for revealing disease mechanisms, pinpointing new therapeutic targets, and eventually moving toward personalized applications.

- Savage and Zhang (2020) — Guide to bioinformatics resources for phosphoproteomics: phosphosite and kinase knowledge bases, prediction of sites and their kinases, and kinase activity inference. Useful for the post-translational modifications subsection. *(review; CC BY 4.0)*

  Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. Savage & Zhang (2020). *Clinical Proteomics*. doi:10.1186/s12014-020-09290-x

  > Mass spectrometry-based phosphoproteomics is becoming an essential methodology for the study of global cellular signaling. Numerous bioinformatics resources are available to facilitate the translation of phosphopeptide identification and quantification results into novel biological and clinical insights, a critical step in phosphoproteomics data analysis. These resources include knowledge bases of kinases and phosphatases, phosphorylation sites, kinase inhibitors, and sequence variants affecting kinase function, and bioinformatics tools that can predict phosphorylation sites in addition to the kinase that phosphorylates them, infer kinase activity, and predict the effect of mutations on kinase signaling. However, these resources exist in silos and it is challenging to select among multiple resources with similar functions. Therefore, we put together a comprehensive collection of resources related to phosphoproteomics data interpretation, compared the use of tools with similar functions, and assessed the usability from the standpoint of typical biologists or clinicians. Overall, tools could be improved by standardization of enzyme names, flexibility of data input and output format, consistent maintenance, and detailed manuals.

- Masson et al. (2019) — Community recommendations that also explain the principles of HDX-MS, how the data are analysed and how differences between states are judged statistically. A good entry point to structural mass spectrometry. *(perspective; CC BY 4.0)*

  Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. Masson et al. (2019). *Nature Methods*. doi:10.1038/s41592-019-0459-y

  > Hydrogen deuterium exchange mass spectrometry (HDX-MS) is a powerful biophysical technique being increasingly applied to a wide variety of problems. As the HDX-MS community continues to grow, adoption of best practices in data collection, analysis, presentation and interpretation will greatly enhance the accessibility of this technique to nonspecialists. Here we provide recommendations arising from community discussions emerging out of the first International Conference on Hydrogen-Exchange Mass

Spectrometry (IC-HDX; 2017). It is meant to represent both a consensus viewpoint and an opportunity to stimulate further additions and refinements as the field advances.

D.8.5 Proteomics without mass spectrometry

- Edfors et al. (2018) — Applies the 'pillars' of antibody validation (genetic, orthogonal, independent antibodies, tagged proteins, capture MS) to thousands of Human Protein Atlas antibodies. Explains why antibody-based proteomics needs systematic validation; it is a primary study rather than a review. *(review; CC BY 4.0)*

  Enhanced validation of antibodies for research applications. Edfors et al. (2018). *Nature Communications*. doi:10.1038/s41467-018-06642-y

  > There is a need for standardized validation methods for antibody specificity and selectivity. Recently, five alternative validation pillars were proposed to explore the specificity of research antibodies using methods with no need for prior knowledge about the protein target. Here, we show that these principles can be used in a streamlined manner for enhanced validation of research antibodies in Western blot applications. More than 6,000 antibodies were validated with at least one of these strategies involving orthogonal methods, genetic knockdown, recombinant expression, independent antibodies, and capture mass spectrometry analysis. The results show a path forward for efforts to validate antibodies in an application-specific manner suitable for both providers and users.

- Eldjarn et al. (2023) — Compares the two main affinity platforms for plasma proteomics, SomaScan (aptamers) and Olink (proximity extension assay), in the same individuals using precision, correlation and genetic associations. A primary study, but the most direct explanation of how the two platforms differ. *(review; CC BY 4.0)*

  Large-scale plasma proteomics comparisons through genetics and disease associations. Eldjarn et al. (2023). *Nature*. doi:10.1038/s41586-023-06563-x

  > High-throughput proteomics platforms measuring thousands of proteins in plasma combined with genomic and phenotypic information have the power to bridge the gap between the genome and diseases. Here we performed association studies of Olink Explore 3072 data generated by the UK Biobank Pharma Proteomics Project 1 on plasma samples from more than 50,000 UK Biobank participants with phenotypic and genotypic data, stratifying on British or Irish, African and South Asian ancestries. We compared the results with those of a SomaScan v4 study on plasma from 36,000 Icelandic people 2 , for 1,514 of whom Olink data were also available. We found modest correlation between the two platforms. Although cis protein quantitative trait loci were detected for a similar absolute number of assays on the two platforms (2,101 on Olink versus 2,120 on SomaScan), the proportion of assays with such supporting evidence for assay performance was higher on the Olink platform (72% versus 43%). A considerable number of proteins had genomic associations that differed between the platforms. We provide examples where differences between platforms may influence conclusions drawn from the integration of protein levels with the study of diseases. We demonstrate how leveraging the diverse ancestries of participants in the UK Biobank helps to detect novel associations and refine genomic location. Our results show the value of the information provided by the two most commonly used high-throughput proteomics platforms and demonstrate the differences between them that at times provides useful complementarity.

- Timp and Timp (2020) — Surveys alternatives to mass spectrometry for identifying and sequencing single protein molecules, such as nanopores and fluorescence-based methods, and their prospects. Background for the subsection on protein sequencing. *(review; CC BY 4.0)*

  Beyond mass spectrometry, the next step in proteomics. Timp & Timp (2020). *Science Advances*. doi:10.1126/sciadv.aax8978

  > Proteins can be the root cause of a disease, and they can be used to cure it. The need to identify these critical actors was recognized early (1951) by Sanger; the first biopolymer sequenced was a peptide, insulin. With the advent of scalable, single-molecule DNA sequencing, genomics and transcriptomics have since propelled medicine through improved sensitivity and lower costs, but proteomics has lagged behind. Currently, proteomics relies mainly on mass spectrometry (MS), but instead of truly sequencing, it classifies a protein and typically requires about a billion copies of a protein to do it. Here, we offer a

survey that illuminates a few alternatives with the brightest prospects for identifying whole proteins and displacing MS for sequencing them. These alternatives all boast sensitivity superior to MS and promise to be scalable and seem to be adaptable to bioinformatics tools for calling the sequence of amino acids that constitute a protein.

D.8.6 Repositories and standards

- Perez-Riverol et al. (2015) — Survey of proteomics repositories and knowledge bases, including PRIDE, PeptideAtlas and ProteomeXchange, and of how their data can be reused. Somewhat dated in details but a good map of the landscape. *(review; CC BY 4.0)*

  Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Perez-Riverol et al. (2015). *PROTEOMICS*. doi:10.1002/pmic.201400302

  > Compared to other data-intensive disciplines such as genomics, public deposition and storage of MS-based proteomics, data are still less developed due to, among other reasons, the inherent complexity of the data and the variety of data types and experimental workflows. In order to address this need, several public repositories for MS proteomics experiments have been developed, each with different purposes in mind. The most established resources are the Global Proteome Machine Database (GPMDB), PeptideAtlas, and the PRIDE database. Additionally, there are other useful (in many cases recently developed) resources such as ProteomicsDB, Mass Spectrometry Interactive Virtual Environment (MassIVE), Chorus, MaxQB, PeptideAtlas SRM Experiment Library (PASSEL), Model Organism Protein Expression Database (MOPED), and the Human Proteinpedia. In addition, the ProteomeXchange consortium has been recently developed to enable better integration of public repositories and the coordinated sharing of proteomics information, maximizing its benefit to the scientific community. Here, we will review each of the major proteomics resources independently and some tools that enable the integration, mining and reuse of the data. We will also discuss some of the major challenges and current pitfalls in the integration and sharing of the data.

- Deutsch et al. (2023) — Describes the HUPO-PSI standards (mzML, mzIdentML, mzTab, Universal Spectrum Identifier, SDRF-Proteomics and others) and why they exist. Directly covers the standards part of the section. *(review; CC BY 4.0)*

  Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work. Deutsch et al. (2023). *Journal of Proteome Research*. doi:10.1021/acs.jproteome.2c00637

  > The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) has been successfully developing guidelines, data formats, and controlled vocabularies (CVs) for the proteomics community and other fields supported by mass spectrometry since its inception 20 years ago. Here we describe the general operation of the PSI, including its leadership, working groups, yearly workshops, and the document process by which proposals are thoroughly and publicly reviewed in order to be ratified as PSI standards. We briefly describe the current state of the many existing PSI standards, some of which remain the same as when originally developed, some of which have undergone subsequent revisions, and some of which have become obsolete. Then the set of proposals currently being developed are described, with an open call to the community for participation in the forging of the next generation of standards. Finally, we describe some synergies and collaborations with other organizations and look to the future in how the PSI will continue to promote the open sharing of data and thus accelerate the progress of the field of proteomics.

- Perez-Riverol et al. (2019) — Analyses how public proteomics and other omics data sets are downloaded and reused. Illustrates why deposition in repositories such as PRIDE matters. *(perspective; CC BY 4.0)*

  Quantifying the impact of public omics data. Perez-Riverol et al. (2019). *Nature Communications*. doi:10.1038/s41467-019-11461-w

  > The amount of omics data in the public domain is increasing every year. Modern science has become a data-intensive discipline. Innovative solutions for data management, data sharing, and for discovering novel datasets are therefore increasingly required. In 2016, we released the first version of the Omics Discovery Index (OmicsDI) as a light-weight system to aggregate datasets across multiple public omics data resources. OmicsDI aggregates genomics, transcriptomics, proteomics, metabolomics and multiomics

datasets, as well as computational models of biological processes. Here, we propose a set of novel metrics to quantify the attention and impact of biomedical datasets. A complete framework (now integrated into OmicsDI) has been implemented in order to provide and evaluate those metrics. Finally, we propose a set of recommendations for authors, journals and data resources to promote an optimal quantification of the impact of datasets.

The following open-access paper(s) are reproduced under their original licences.

proteomes

MDPI

Review

A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field

Emmalyn J. Dupree [1,†], Madhuri Jayathirtha [1,†], Hannah Yorkey [1], Marius Mihasan [2,*], Brindusa Alina Petre [3,4,*] and Costel C. Darie [1,*]

1 Biochemistry & Proteomics Group, Department of Chemistry and Biomolecular Science, Clarkson University, 8 Clarkson Avenue, Potsdam, NY 13699-5810, USA; dupreeej@clarkson.edu (E.J.D.); jayathm@clarkson.edu (M.J.); yorkeyh@clarkson.edu (H.Y.)
2 Laboratory of Biochemistry & Molecular Biology, Faculty of Biology, "Alexandru Ioan Cuza" University of Iasi, 700506 Iasi, Romania
3 Laboratory of Biochemistry, Department of Chemistry, "Alexandru Ioan Cuza" University of Iasi, 700506 Iasi, Romania
4 Center for Fundamental Research and Experimental Development in Translation Medicine–TRANSCEND, Regional Institute of Oncology, 700506 Iasi, Romania
* Correspondence: marius.mihasan@uaic.ro (M.M.); brindusa.petre@uaic.ro (B.A.P.); cdarie@clarkson.edu (C.C.D.)
† These authors contributed equally to this work.

Received: 1 June 2020; Accepted: 1 July 2020; Published: 6 July 2020



Abstract: Proteomics is the field of study that includes the analysis of proteins, from either a basic science prospective or a clinical one. Proteins can be investigated for their abundance, variety of proteoforms due to post-translational modifications (PTMs), and their stable or transient protein–protein interactions. This can be especially beneficial in the clinical setting when studying proteins involved in different diseases and conditions. Here, we aim to describe a bottom-up proteomics workflow from sample preparation to data analysis, including all of its benefits and pitfalls. We also describe potential improvements in this type of proteomics workflow for the future.

Keywords: mass spectrometry; proteomics; protein identification; protein characterization

1. Proteomics

The term proteome refers to all proteins that are produced or modified by an organism (e.g., human [1], animal [2], plant [3], bacteria [4]) or living system (e.g., organ, cell culture, complex community from an environmental sample)). The term "proteome" and the first dedicated proteomics laboratory were introduced in 1994 by Wilkins et al. to describe proteins as a complement to genomic data [5]. However, the "whole" proteome of a particular cell, tissue, organ, or organism is yet to be identified. This is particularly difficult due to the vast variety of proteins and their isoproteins/proteoforms/protein species, which are expressed at different levels—from very abundant proteins, such as actin, to less abundant ones, such as transcription factors—in different cells, tissues, or organs. The variety of post-translational modifications (PTMs) in proteins, which may be stable or transient, is responsible for the vast number of proteoforms, which is an obstacle in most proteomics experiments. This, corroborated with the multiple limitations of proteomics methods, makes the quest to identify the proteome of any given organism a difficult one [6].

The proteomics field consists of a wide range of methodology, which has been largely driven by the modern development of involved technology. The concept of global protein analysis as a complete atlas of human proteins was proposed over 50 years ago [7]; however, proteomics research did not start until the mid-1990s. The beginning of proteomics research was sparked due to parallel

developments in four areas: (i) two-dimensional gel electrophoresis (2D-PAGE) evolving into a robust method to rapidly separate proteins contained in biological complex mixtures [8]; (ii) the continuous development of mass spectrometry methods for accurate mass and chemical structure analysis [9,10]; (iii) the constant production of large-scale genome research and enormous number of peptide/protein sequences catalogued in several databases [11]; and (iv) the development of novel bioinformatics tools to simplify the analysis of large volumes of MS data aiming towards identifying intact proteins and their functional or pathological PTMs [12,13].

The field of proteomics comprises a bioanalytical discipline that performs large-scale studies on proteins [14] that may be rooted from a basic science perspective or a clinical one, i.e., proteins that are associated with a broad range of diseases and conditions by means of their altered expression levels and/or PTMs. In addition to fundamental protein research or development of proteomics-related instrumentation, the detailed analysis of the proteome for a specific type of cell (e.g., tumor, blood, or tissue) has the potential to permit the discovery of new protein biomarkers aimed towards diagnostic purposes and novel drug discovery [15,16]. Currently, the knowledge provided by proteomic research adds greatly to the genetic information generated from all genomics studies. The combination of proteomics and genomics research has the potential to play a major role in future biomedical studies and to impact the development of next generation diagnostic and therapeutic approaches.

Mass spectrometry (MS)-based proteomics has led to the possibility of characterizing and quantifying the protein profile of biological specimens, as well as the possibility to discover their complex interactions involved in various specific pathologies. For example, various proteomic approaches combined with genomic analysis have been used in cancer research for obtaining more information about the molecular basis of tumor genesis and the development of more effective anticancer therapies [17–19].

A typical bottom-up proteomics workflow (Figure 1) consists of several major steps: (i) isolation of the protein mixture from the studied biological sample, followed by (ii) quantification of isolated proteins concentrations (e.g., Bradford assay), then (iii) fractionation of proteins by gel electrophoresis or liquid chromatography methods. After fractionation, (iv) the proteins are proteolytically cleaved by enzymes (usually trypsin); followed by (v) a mass spectrometric measurement of the resulting peptides and (vi) a database search for protein identification.

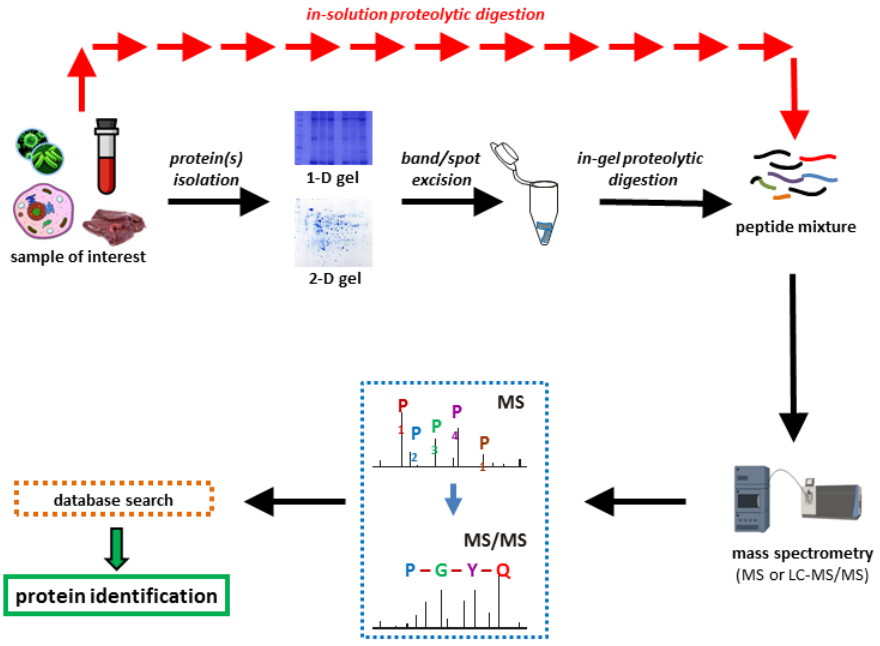


Figure 1. Workflow in a bottom-up proteomics experiment.

The purpose of this review is to briefly introduce the bottom-up proteomic approach from a liquid chromatography–mass spectrometry perspective, to discuss its multidisciplinary development, its strengths and weaknesses, and to identify potential areas of improvement in the current methodology. These potential improvements will be relevant to the current problems in society, including biomedical, clinical, or environmental concerns.

2. Bottom-Up Proteomics

Most proteomic analyses use proteases to digest proteins into peptides with a predictable terminus. Peptides are then analyzed in an MS/MS instrument and their mass to charge ratio and predicted sequence is used to infer information about the proteins in the sample. All experimental set-ups starting from the analysis of peptides from complete protein digests and using a protein database to characterize the open-reading frame this peptide originated from, are grouped under the term "bottom-up" proteomics.

Bottom-up proteomics utilizes the advantages that peptides have over proteins: peptides are more easily separated by reversed-phase liquid chromatography (RPLC), ionize well [20], and fragment in a more predictable manner. This translates into a robust methodology that enables high-throughput analysis, allowing for identification and quantification of thousands of proteins from complex lysates [21]. Today, bottom-up approaches using data-dependent acquisition (DDA) workflows are the core technologies in proteomics. Also termed shot-gun proteomics, these straightforward workflows generate large lists of protein identifications and were used for solving most of the available, complex, full proteomes available today, including the first drafts of the human proteome [22,23]. New data acquisition strategies such as selected reaction monitoring (SRM) further increase the quantification accuracy and reproducibility of bottom-up proteomics studies [24]. Data independent acquisition (DIA) is evolving to become a major driving methodology of the future. It attempts to combine a high number of protein identifications with consistency and accuracy in quantification of protein levels [25].

The hallmark of bottom-up proteomics is the extensive use of protease digestions, which has its drawbacks. Trypsin is the golden standard when it comes to shot-gun approaches and is being used for approximatively 96% of the deposited data sets in the Global Proteome Machine Database [26]. Trypsin is a very efficient protease with high catalytic activity that generates peptides with a basic arginine or lysine at the C-termini, which is ideal for collision-induced dissociation (CID) tandem mass spectrometry analysis [27]. Although efficient, when using trypsin, 56% of all generated peptides are ≤ 6 residues and thus too small to be identified by MS [28]. Moreover, only a fraction of these peptides yields useful fragmentation ladders [29]. The extensive usage of trypsin also leads to analytical tools that are optimized based on tryptic peptide properties. Furthermore, protein identifications are usually inferred by a limited number of typtic peptides. Hence, current bottom-up proteomics offers a rather biased and restricted coverage of the full proteome in a given sample, almost like "tunnel vision" of the proteome [28].

The limited sequence information inferred from small peptides is more often suffice to assign proteins clusters, and not always enough to identify proteoforms. Although possible through the measurement of mass shifts, protein isoform and PTM identification without prior knowledge is extremely limited in bottom-up proteomics.

3. Sample Preparation

As the proteomics field attempts to perform a comprehensive analysis of all the proteins in a given sample [30], sample preparation is the first and most important step in any experimental endeavor. If this step fails to perform, the sensitivity and throughput of the downstream steps are rendered pointless.

Genomics and transcriptomics, the other two major high-throughput -omics fields, benefit from the fact that nucleic acids are relatively uniform in terms of physiochemical properties. As such,

"universal" standard methods and buffers are available that can be used to isolate nucleic acids of any size from almost any sample. No such thing exists in proteomics. The functional groups of the amino acid sidechains allow for such a great diversity in terms of protein charge and hydrophobicity, that a universal buffer will most probably never be available [31].

Nucleic acid-based technologies also benefit from the power of the PCR reaction, which allows the production of large quantities of target DNA in a single, reliable, and easy to do step; however, the proteomics field has nothing of the sort. Biological protein mixtures are characterized by a great diversity of proteins in chemistry, shape, and size, and, most importantly, by their range in abundance. Some proteins can be found in quantities as low as one protein per cell, while others up to several million per cell [32]. Despite the great selectivity and sensitivity of mass-spectrometers, no instrument available today can solve such complexity of range in a single measurement. Thereby, various fractionation strategies must be used in order to improve the depth and coverage of proteomic analysis [33].

3.1. The Good

Usually, methodologies for bottom-up proteomic sample preparation include the extraction of proteins from the biological matrix, removal of the nonprotein contaminants such as DNA, sugars, and lipids, removal of the residual salts that may form adducts during ionization, and protein fractionation to reduce sample complexity.

Bottom-up proteomics utilizes unfolded proteins that allow for easy access for proteases since amino acids are more readily available, thus, generating more peptides for MS analysis. In this case, tissue or cell lysis is performed directly in a buffer containing strong denaturants (such as urea or guanidine) and ionic detergents (sodium dodecyl sulfate, SDS, or deoxycholate-SDC). In addition, nonionic zwitterionic detergents (such Triton X-100, NP-40, digitonin, or CHAPS) further help solubilize membrane proteins and can be used when less denaturing conditions are required [34]. Depending on the sample, sometimes protein depletion or enrichment is a crucial step—usually mandatory for serum or plasma samples where albumin represents more than 50% of the proteins. Several methodologies are available for removing high-abundance proteins [35,36] or for enriching cell surface proteins [37], phosphoproteins [38], or various enzyme subclasses [39–41]. The techniques that provide the highest selectivity and sensitivity for the low abundant proteins are affinity chromatography [42] and immunoprecipitation [43], but their applicability depends on the availability of affinity supports or antibodies, as well as the quality of said antibody for the target protein.

For the removal of small molecules that may have originated from the biological sample or were added in the extraction/depletion/enrichment, many methods are available, including dialysis, buffer exchange, size exclusion, protein precipitation, chromatography, or electrophoresis [44,45]. Of all the desalting methods, precipitation with organic solvents (acetone or methanol/chloroform) is the least expensive, simplest, and most scalable option for desalting proteins prior to MS analysis [20].

Two forms of gel electrophoresis, SDS-PAGE and 2D-PAGE, have been extensively used in proteomics analysis as they solve two problems simultaneously—they remove salts, detergent, and other small molecules from the sample and they fractionate the proteins based on apparent molecular mass (SDS-PAGE) or charge and apparent molecular mass (2D-PAGE). SDS-PAGE has the advantage that it is inexpensive, straightforward, and highly reliable. Its separating resolution is higher than size exclusion chromatography, but not as high as 2D-PAGE. 2D-PAGE has a high resolution and, with the introduction of immobilized pH gradient (IPG) strips, its reproducibility has improved considerably, becoming the method of choice for most proteoform studies [46,47]. The development of PTM-specific stains for phosphoproteins (e.g., Pro-Q Diamond) and glycoproteins (e.g., Pro-Q Emerald, Dansylhydrazine) further increased the applicability of 2D-PAGE [48]. 2D gels are very useful for de novo sequencing of proteins from organisms with no genome sequences available and can complement the identification of protein isoforms and modified proteins [49].

3.2. The Bad

When preparing a sample for bottom-up proteomics analysis, many factors can cause the experiment to fail. Some examples include incomplete lysis, incomplete solubilization of the proteins, and contamination during sample handling with compounds that interfere with the downstream analysis steps, including MS. Some of the most troublesome compounds are detergents—both the detergents used for sample solubilization and those used for cleaning laboratory glassware. Detergent-contaminated samples and foul autosampler needles and LC piping reduce column capacity and performance and have poor ionization in both ESI and MALDI [20]. Triton X-100, Tween, or NP-40 contain polyethylene glycol (PEG) chains that elute throughout the LC and overwhelm the MS detector. SDS can cause complete signal suppression at levels as low as 0.01% [50]. Not all detergents are the same—some are considered safer for mass spectrometry, such as N-octyl-beta-glucoside and octylthioglucoside [51]. Moreover, several mass-spectrometry-compatible detergents are available, such as ProteaseMax (Promega), Rapigest (Waters), PPS Silent Surfactant (Expedeon), or Progenta (Protea), which degrade with heat and in the low pH of the acid containing LC-MS buffers [20]. Compounds leaching from poor quality plastics and fittings are also common contaminants. Phthalates, for example, are known to ionize well and overwhelm the mass spectrometer. An extensive list of contaminates in mass spectrometry is provided by Keller, 2008 [52].

In many proteomic experiments, the immunoaffinity methods are used to deplete the most abundant proteins. For example, there are many methods available that deplete albumin from human serum. Although these methods allow for the identification of a larger number of proteins due to the decrease in signal suppression, there may be problems with the nonselective loss of other proteins [53].

Keratins from skin and dust are another common contaminant, which takes up detector time, clutters the raw data, and reduces the number of useful spectra recorded. Keratins are commonly introduced when casting and handling gels and gel pieces [54]. It is thereby imperative to wear gloves and scrupulously wipe down surfaces to minimize this contamination. Other frequent proteinaceous contaminants are lysozyme, DNase, and RNase introduced in the cell lysis step, bovine serum albumin and other weight marker components from electrophoresis, as well as trypsin and other proteases from the digestion step [52]. A significant amount of MS instrument time is spent sequencing peptides from these abundant contaminant proteins and not peptides from the actual sample of interest [55]. Exclusion lists have been put together in an attempt to cleverly remove the burden of these contaminants [55]; however, these lists are useless in certain instances when, for example, keratins are expected to be present in the sample to be analyzed. In this case, it is difficult to distinguish between endogenous and contaminating keratins [56].

SDS-PAGE is an easy, inexpensive method to separate proteins prior to mass spectrometry analysis; however, its separating power is not very high. Although 2D PAGE has better separating power, it is still not uncommon to have multiple proteins in the same spot [57]. Along with this, proteins with a low copy number and hydrophobic proteins are difficult to detect in 2-D electrophoresis [58].

Ultimately, difficulties in sample preparation for bottom-up proteomics analysis are a result of the huge variability of protein properties. There is an overwhelming number of methods and technologies available, each tailored for specific samples and protein groups. With such variety to choose from, one will have a hard time deciding which method to pick and how they should adapt it to fit the specific sample. Moreover, most of the methodologies are long, multistep preparations that are prone to losses, biases, and contaminations, while being extremely time-consuming and labor-intensive. In the end, all of this ultimately contributes to poor reproducibility of the current sample preparation and fractionation methods [59].

3.3. The Future

Ideally, what is needed is a reliable and universal method to extract and fractionate proteins; one that is fast, does not require detergents, requires minimal steps and sample handling, and can be automated. This may be achievable with the development of new approaches that rethink sample

preparation in bottom-up proteomics. Sample Preparation by Easy Extraction and Digestion (SPEED) is a new approach that does what its name suggests—reduces sample preparation to three easy steps: (1) acidification with pure trifluoroacetic acid (TFA) to lyse cells and extract proteins, (2) neutralization with TRIS base, and (3) digestion with proteases to generate peptides [59]. All steps and buffers required for sample preparation can be integrated for a straightforward and possibly automated sample preparation. Depending on the protein digestion mechanism, three types of integrated sample preparation methods are emerging: (1) in solution digestion, (2) immobilized-enzyme-reactor, and (3) on bed digestion methods [60]. Integrated in-solution digestion methods include filter-aided sample preparation (FASP) that repurposes centrifugal ultrafiltration concentrators in order to remove detergents, perform protein cleavage and isolate peptide fractions [61,62]. Derived from FASP, an encapsulated in-StageTip (iST) device has an enclosed tip chamber with an inserted membrane where lysis, denaturation, and alkylation take place [63]. The membrane serves both as a filter and as separation support. NanoPOTS are microfluidic chip devices that use nanodroplet-based processing methods. They allow in-solution digestion and processing of single cells and are reproducible [64]. Immobilized enzymatic reactor (IMER) methods make use of special columns that contain immobilized trypsin and can be directly coupled to LC-MS system for automated protein digestion and online LC-MS/MS analysis [65]. On-bead digestion methods perform reduction, alkylation, and digestion steps on functionalized beads by trapping proteins into a very limited void volume. Suspension trapping (STrap) is a method that uses tips packed with a quartz or glass filter and a hydrophobic C18 layer to perform sample clean-up and protein digestion in one step [66]. Single-Pot Solid-Phase-enhanced Sample Preparation (SP3) uses paramagnetic beads for the same purpose [67] and can be easily automated using a liquid handling robot [68]. These methods can be easily automated and integrated.

4. Conventional HPLC/Modern UHPLC Fractionation

Currently, liquid chromatography coupled with mass spectrometry (LC-MS) is a common and indispensable analytical technique for proteomic investigation. This coupling initiated the development of novel ionization methods and led to a broad range of interfaces aimed at separating various biological complex mixtures [69]. Atmospheric pressure ionization (API) was the first method to directly interface a solution stream with a mass analyzer [70]. Several thermospray units were also introduced as a breakthrough for modern LC-MS [71,72].

4.1. The Good

The main applications of HPLC in proteomics are rooted in the concept that peptides can be separated over a time and buffer gradient. This allows for maximum identification by the mass spectrometer, helps with the problem of ion suppression from coeluting peptides, and helps in the identification of the individual protein sequence. Significant research has been aimed towards developing HPLC methods to enable the resolution and identification of all generated peptides from digested proteins in a given proteome. This is a difficult task for many reasons. For example, a serum proteome may contain up to 20,000 proteins with a concentration dynamic range of 10^{11}, which, by proteolytical digestion, may result in more than 600,000 peptides [73] excluding PTMs. Fractionation of such complex peptide mixtures is a critical aspect of their mass spectrometric identification. The general principle of HPLC fractionation of complex peptides mixtures is based on their interaction with a stationary phase (column) and a mobile phase (solvent gradient elution). The three major HPLC modes developed for peptide fractionation or sequential separation utilize differences in peptide size (Size-Exclusion Chromatography (SEC)), net charge (Ion-Exchange Chromatography (IEX)), or hydrophobicity (Reverse-Phase Chromatography (RP-HPLC)) [74,75].

The well-established proteomic methods approaching structural characterization of proteins (top-down and bottom-up approaches), have led to the use of HPLC methods in a particular mode [76]. For example, SEC [77] was used in top-down approaches to study intact proteins as a favored method for size-based separation. SEC has been mostly selected to analyze antibody–drug conjugates

to determine their purity [78] and to purify recombinantly expressed proteins [79]. Since SEC is considered a low-resolution chromatographic method that requires diluted samples, the SEC technique was employed in a combination of columns with different pore sizes to achieve sufficiently high resolution in the separation of a complex protein mixture with a broad molecular weight range (10–223 kDa) [80].

Moreover, the combination of SEC with RP-HPLC, in a two dimensional (2D) separating platform exceeded a one-dimensional (1D) RP-HPLC experimental run with 4044 more unique proteoforms identified in a sarcomeric protein mixture [80]. This multidimensional HPLC was introduced as a shotgun approach to analyze very complex protein or peptide mixtures without performing gel electrophoresis, but achieving the same separating resolution as bidimensional gel electrophoresis (2D-PAGE) [81]. In the so-called multidimensional protein identification technology (MudPIT) [82,83], a protein mixture was subject to specific enzymatic digestion, usually using trypsin and endoproteinase LysC, and the resulting peptide mixture was separated by strong cation exchange (SCX) and reversed-phase high performance liquid chromatography (RP-HPLC) [84,85]. A combination of HPLC, liquid phase isoelectric focusing, and capillary electrophoresis was reported as a multimodular approach for obtaining a better separation of complex protein mixtures [86]. Recently, the large dynamic range and complexity of human cancer cells was overcome by employing a 2D separation using a high pH and low pH reversed-phase liquid chromatography technique, and 2778 proteoforms from 628 intact proteins were detected [87].

RP-HPLC is used in most bottom-up proteomic experiments for separating proteolytically generated peptides due to its high peak capacity, reproducibility, and robustness [88,89]. Both bottom-up and top-down proteomics approaches are fully dependent on employed separation technologies to: (i) provide large-scale proteome coverage in a given time; (ii) accomplish higher analytical throughput; and (iii) cover a broad dynamic protein concentration range, including trace amounts of distinct proteins. Many proteomic approaches use a IEX/RPLC [90,91] combination where octadecylsilanes (C18) remains the preferred RP ligand and the choice of ionic ligands experimentally depends on the class of peptides to be enriched and/or fractionated.

The development of an HPLC column with an "ideal and perfect" peptide/protein separation capability is an important, continuous research objective. The main types of analytical columns currently used in proteomics research present different characteristics with regard to the material composition and particle size packed into it, and the length and diameter [92]. Remarkable separation of complex peptide mixtures in proteomic studies was achieved using nano-LC/UPLC and capillary columns containing packed alkyl bonded C8 or C18 [93] and silica-based monolithic capillaries [94,95]. In proteomic analysis, the ability to handle very small amounts of biological material is crucial. Miniaturized HPLC separation systems have been developed by using fused silica capillary columns [96,97] or chip-based devices [98–100]. Monolithic capillary columns (and other columns, such as C8 or C18 columns) can be produced in the laboratory without the need of expensive media, packing solvents, and high-pressure packing instrumentation with the proper training [101]. Even if reproducibility issues of published protocols to produce such columns were reported, they remain cost-efficient and depend on the researcher's capability to pack it perfectly.

Separation time is an important factor in nanoflow LC/tandem MS technologies, which has been shown by Mann et al. [102] where the results on the yeast proteome were compared using different mobile-phase gradient running times, identifying 5806 peptides in 140 min and 13,682 peptides in 480 min. Furthermore, improved fractionation in chromatography was shown as a straightforward approach to separate coeluting peptides and the pH of the mobile phase was shown to have an important impact on the retention, selectivity, and sensitivity of the separation [103].

In 1997, MacNair introduced a hybrid stationary phase for ultrahigh-pressure reversed-phase liquid chromatography (UHPLC) for rapid separation [104]. The difference between UHPLC and traditional HPLC is that UHPLC uses smaller diameter packing material in the column and higher pump pressures than HPLC. In UHPLC separation, the size of the particles is usually approximately

1.7 µm in diameter, versus the 3–5 µm diameter of HPLC column particles and requires a pressure of up to 1000 bar, versus the 50–600 bar range that is typical in HPLC instruments [105]. UHPLC has raised the level of performance of the separation with significant gain in resolution, speed, and sensitivity. The use of UHPLC doubled the peak capacity and increased the separation speed over nine-fold and the sensitivity by three- to five-fold, as compared with an HPLC run [106]. The problematic thermal effect occurring in the column by using high pressure in UHPLC system was overcome by their narrower diameters. However, as column diameters decrease, technical problems such as (i) sealing the LC system to resist leakage, (ii) very high backpressures at high flow rates, and (iii) contamination were reported. In terms of primordial advantages of UHPLC over HPLC, there is the faster speed and less solvent consumption, but its evident disadvantage is the price [107].

4.2. The Bad

A problem that remains in part unresolved by liquid chromatography is membrane protein solubility, which requires the use of special solubilization agents such as acidic solvents, detergents, and/or chaotropes that need to be diluted or even removed prior to proteolytical digestion or mass spectrometric analysis [108].

To optimize fractionation protocols using multidimensional separation, several factors need to be considered, including the impact of ionic strength, buffer capacity, pH response on the retention time behavior, and peak shape of proteins or peptides [109]. Moreover, when using multiple SCX /HPLC separation, the elution of the peptides from the SCX column is not precise and the same peptide will appear in several of the subsequent HPLC runs, reducing the amount of that peptide and thus decreasing the sensitivity. In samples that are too complex/concentrated, the first few buffer injections will contain the majority of singly charged peptides and di/triprotonated peptides are usually not trapped [110]. The major obstacle for liquid chromatography in using multiple parallel columns is adjusting the hardware and/or software to work in a synchronic manner. The price of the columns and their stability is not to be neglected. Future contributions in column technology require important developments so that LC is not restricted to the conventional single-column fractionation methods.

Broadly, pharmaceutical and clinical laboratories seem to be willing to sacrifice resolution to gain analysis speed. This sacrifice may not be acceptable in fundamental research when analyzing complex peptide/protein mixtures to provide data for developing novel drugs or when identifying new biomarkers with diagnostic purposes. HPLC/UHPLC will continue to be a pivotal analytical technique that, in combination with high resolution mass spectrometry, will further raise the level of performance with significant increases in resolution, speed, and sensitivity required for elucidation of complex biological of proteomes [111].

4.3. The Future

HPLC/UHPLC-MS will remain very useful in "omics" sciences and reference methods will be developed, but the transition to various clinical applications is more desired. In recent years, immunoassays have been progressively replaced by HPLC-MS due to higher sensitivity, significantly less false positives, and the reduced costs of used reagents because multiple analytes can be measured simultaneously [112]. On the other hand, some drawbacks of HPLC-MS analysis include the high cost of instrumentation and the necessity for the initial method development and rigorous validation. As of today, universal chromatographic methods are not available, and each laboratory has to develop their own methods critically depending on their available infrastructure. The improvement of proteome coverage by first reducing sample complexity via chromatography will be a great challenge to researchers. One important requirement that gets increasingly supported by the literature is the fact that the current column hardware is no longer adequate to maintain the very high efficiencies and small peak volumes produced by the high-quality particles (continually smaller) and high-quality packing procedures. The proteomics field has also been utilizing very slow microflow rates in order to increase sensitivity, which cuts down on buffer costs, but also increases analysis time [113].

As several method parameters, such as pH, temperature, buffer concentration, and gradient time are varied simultaneously, a revolutionary software that predicts chromatograms based on HPLC method development is desired to determine the behavior of the separation. This is necessary in order to optimize complex sample separation and to economize resources spent developing and running wet lab activities. With the help of bioinformatics developments, in silico chromatography will be of great help in the near future to predict LC experimental flows for avoiding time, solvent, and sample consumption normally used in optimizing protocols for complex biological peptide/protein mixtures. Recently, a predictive algorithm of peptide and protein retention times in reversed-phase chromatography [114] was reported as a complementary experimental tool for proteomics.

5. MS Analysis: Instrumentation

MS-based proteomics is one of the chosen methods for complex protein sample analysis. It has established itself as a superior technology for complete characterization of proteins [29]. It is widely used for sequence analysis, protein–protein interactions, and identifying PTMs [29]. In general, a mass spectrometer consists of an ion source to ionize the analytes, a mass analyzer to measure the mass to charge ratio (m/z) of the analytes, and a detector that detects the number of ions at each m/z value. Electrospray ionization (ESI) is a commonly used technique to ionize peptides or proteins for MS analysis [115]. It ionizes a liquid solution of sample and hence can be coupled to liquid chromatography for separation [29,115]. There are four different type of mass analyzers used in proteomics: ion trap, quadrupole, time of flight (TOF), and Fourier transform ion cyclotron (FT-MS). These analyzers are the key to maintain sensitivity, mass accuracy, resolution, and to generate information rich ion mass spectra (MS/MS spectra) from peptide fragments [116]. They can be used individually in instruments or combined with each other to take advantage of the strengths of each [29,116].

It is also necessary for peptides be broken up further before being analyzed by the mass spectrometer, by a process known as dissociation. Different dissociation techniques exist for mass spectrometers, including (1) collision-induced dissociation (CID), (2) electron-capture dissociation (ECD), (3) electron-transfer dissociation (ETD), and (4) higher-energy collisional dissociation (HCD).

5.1. The Good

There is a diverse range of MS instruments that cover many possible applications in proteomics. Recent developments in the instrument design have led to new ion activation techniques as well as allowing for lower limits of detection. It has also increased the ability of tandem mass spectrometry for peptide and protein structure elucidation by improving the understanding of gas-phase ion chemistry [117]. In addition, the dynamic range of instruments allows one to optimize LC-MS and LC-MS/MS methods such as Data Dependent Analysis (DDA), Data Independent Analysis (DIA), Selected Reaction Monitoring (SRM), and Parallel Reaction Monitoring (PRM).

DDA is a common data acquisition strategy that selects the most abundant precursor ions for MS/MS analysis [118]. It takes the selection of peptide signals forward for fragmentation and matches them to a predefined database. The method allows for minimal selection of redundant peptide precursors [119]. The semirandom peptide sampling phenomenon in a DDA method has shown to increase the rates at which new peptides are identified, especially during replicate analysis. However, the reproducibility of the low abundant peptides between the runs, remain a challenge in the random sampling situation [118,120].

DIA is a method in which all the peptides within a defined m/z frame are subjected to fragmentation [121]. This method allows for accurate peptide quantification without being restricted to profiling only the predefined peptides of interest. In addition, DIA has the potential to overcome the random sampling problem and to reproduce and quantify low abundant peptides [121]. DIA offers several advantages over DDA for characterizing complex proteins. Unlike DDA, which sequentially detects, selects, and analyzes the individual ions, DIA systematically parallelizes the fragmentation of detectable ions within an m/z window, regardless of their intensity, thus providing a broader dynamic

range of detected signals, improved reproducibility for identification, better accuracy and sensitivity for quantification, and enhanced protein coverage [122].

SRM is an MS-based technique for quantitative analyses. It is a nonscanning technique where selectivity is increased through fragmentation [123]. This method has the benefit of being able to control error rates in discovery proteomic experiments and can efficiently generate specific and quantitative assays for a large number of proteins and their PTMs by being both cost and time efficient [124]. It serves as a benchmark to all time segment methods [125]. PRM is another targeted method of quantitation performed using high-resolution mass spectrometers such as quadrupole-Orbitrap (q-OT). The development and application of higher-energy collisional dissociation (HCD) fragmentation enables MS/MS spectra to be acquired in the Orbitrap analyzer with high mass accuracy and high resolution. HCD is a beam-type collisional dissociation similar to the dissociation achieved in QQQ as well as QTOF mass spectrometers. An advantage of using q-OT is that both the discovery and targeted experiments can be performed on the same instrument, and it is convenient to transfer instrumental parameters such as collision energy, retention time, quadrupole isolation window, etc. [126]. This approach enables the acquisition of full MS/MS spectra of a targeted peptide with high mass accuracy and resolution, thus quantifying highly specific proteins. Similar to SRM, PRM can also validate the abundance of proteins and their PTMs [126]. While SRM and PRM are comparable, PRM is the most suitable for an attomole-level detection and quantification of multiple proteins in a complex sample [127]. Its simple and straight forward acquisition method in addition to its high selectivity and specificity of data acquired from high resolution and high mass accuracy makes it a powerful quantitation method [128]. Thus, one can choose from one of the previously mentioned methods based on their specific needs and optimize these methods in the instrument in order to acquire data in a short period of time with accurate peptide quantitation.

Recently, the combination of the linear ion trap with the Orbitrap analyzer has benefitted the advances in high resolution MS greatly [129]. It can achieve high mass resolutions within a fraction of a second, which is important in both qualitative and quantitative applications. Moreover, it allows for the identification and quantification of a compound even in the presence of a background ion that has a nominally identical mass [129]. Complementing the ion trap and Orbitrap combination, the quadrupole mass filter was coupled to an Orbitrap analyzer known as the "Q Exactive" instrument, which features fast collision-induced dissociation peptide fragmentation with high energy because of parallel filling and detection modes, and features high ion currents because of an S-lens. This combination aids multiplexed operation at the MS and tandem MS levels, enabling joint analysis of HCD fragment ions in the Orbitrap analyzer and fragmentation of different precursor masses by the quadrupole analyzer, overall making Q Exactive an exciting instrument for proteomics [129]. SWATH-MS is a DIA–LC-MS technique that has become prominent for quantitative proteomics mainly due to its high quantitation accuracy, increased peptide coverage, generation of a digital map, and excellent reproducibility. In addition to that, it also permits qualitative analyses and allows for small molecule applications such as forensic analysis and the identification of metabolites and metabolomics [130]. MS instruments are usually compared to Nuclear Magnetic Resonance (NMR) instruments because they are expensive and require regular maintenance and troubleshooting. However, the robustness, the relative simplicity of sample preparation steps, high-throughput analysis, and sensitivity of MS remains unmatched [131,132].

Ion mobility mass spectrometry (IM-MS) uses electric fields to drag analytes through a buffer gas, separating the analytes whilst providing structural information, and has had many improvements in recent years [133]. The extra dimension that ion mobility provides increases the peak capacity in LC-MS workflows [134]. It is most common to couple a TOF mass analyzer with ion mobility; however, other analyzers can be used.

5.2. The Bad

Mass spectrometers are very expensive, delicate, and require a significant amount troubleshooting and maintenance. Although they are known to produce extensive information on proteins, the abundance of data can give rise to negatives, false positives, and unassigned spectra. There is a dynamic range of instruments to cover all possible applications; however, there is not one specific instrument that can perform all kinds of experiments [132]. MS detection oscillates with the concentration range of 10^4 and 10^5 and sometimes even 10^7. Overall, the complexity, dynamic range of biological samples, and low abundance of disease-specific biomarkers remains a major challenge for proteomic biomarker discovery, and there is no MS instrument that can simultaneously address these challenges efficiently [135]. In addition, complete characterization of the proteome both quantitatively and qualitatively remains a challenge [136]. These problems can be remedied by reducing the sample complexity before introduction into the MS. Another major challenge is the analysis of a large number of samples in clinical studies or discovery proteomics where there is often experimental variability among the clinical samples. Additionally, the impact of single nucleotide polymorphisms (SNPs) on proteome analysis has still not been fully investigated [136]. Proteomics experiments for hundreds of samples are expensive and time consuming [132]. Very few proteomics techniques allow for high throughput analysis while simultaneously maintaining the sensitivity and robustness. DIA/SWATH, as mentioned above, is an alternative for proteomic analysis of clinical samples on a large scale [132,137].

5.3. The Future

Although mass spectrometers have the upper hand concerning their high throughput analysis and robustness, there is a need for MS with better sensitivity, resolution, and accuracy for analyzing samples such as proteins in the blood plasma. This will further allow the instrument to detect low-abundance proteins and their PTMs or interacting partners [132]. In addition, conventional MS has trouble resolving multiple charge states for species such as protein complexes of several hundred kilodaltons, especially for heterogenous samples such as heavily glycosylated proteins. Instruments such as charge detection mass spectrometry (CDMS), a variant of MS can help overcome such complexities by allowing the detection of m/z and charge states, making it easy to determine the mass of ions in a sample [138]. Advances in MS methods will aid in improving the chances of biomarker discovery for MS-based proteomics. Bottom-up proteomics is the workhorse for proteomic analysis. The middle and top-down approaches must advance in order to completely characterize the protein isoforms as well as the PTMs [136]. Overall, continued improvements are needed in both MS methods and technology in order to overcome the aforementioned challenges.

6. Analysis of Mass Spectrometry Data

Analysis of MS data proves to be a limiting factor in many proteomics experiments. Although direct analysis of the raw data can be very useful, this practice can be tedious and time-consuming. There are many different software pipelines available for a more high-throughput and time-efficient approach; however, these programs have a great number of limitations. Although there are still many obstacles to overcome, in recent years, the limitations in proteomic data analysis have lessened, and high-throughput data analysis continues to improve tremendously.

It is important to understand the basics of peptide fragmentation before one can grasp a firm understanding of the current problems in proteomic data analysis, since most problems arise from lacking complete sequence information for many proteins. Peptides produced by enzymatic (i.e., trypsin) digestion will have between 5 and 20 amino acids. During collision-induced dissociation (CID) MS/MS fragmentation, the peptide is fragmented into many smaller fragment ions. In CID, these fragments are called y ions and are numbered from y1 (the C-terminal amino acid) to y(n), where (n) is the maximum number of amino acids in the peptide. A second type of ion in CID is the b ion. These ions start from the N-terminus, and end at the C terminus of the peptide. Both y

and b ions are produced through fragmentation at the peptidic bond—the weakest bond in peptides (Figure 2). Additional ions are also produced in CID and observed in MS/MS: a and c (and b) ions from the N-terminus of the peptide and x and z (and y) ions from the C-terminus of the peptide (Figure 2). Predictability of peptide fragmentation (and production of the y, b, and a ions) allows for identification of peptide sequences using the de novo sequencing and in almost any database search algorithm. Different dissociation techniques produce different types of fragment ions. For example, ECD is complementary to CID as it provides more extensive sequence coverage. Disulfide bonds are preferentially cleaved in this method, and PTMs tend to stay intact; however, this method is specific to FTICR MS instruments [139]. Similarly, ETD fragments in the same type of way, creating longer c and z-type ions and preserving PTMs, but instead uses an RF quadrupole ion trapping device [140]. If used together, these dissociation techniques can give complementary information.

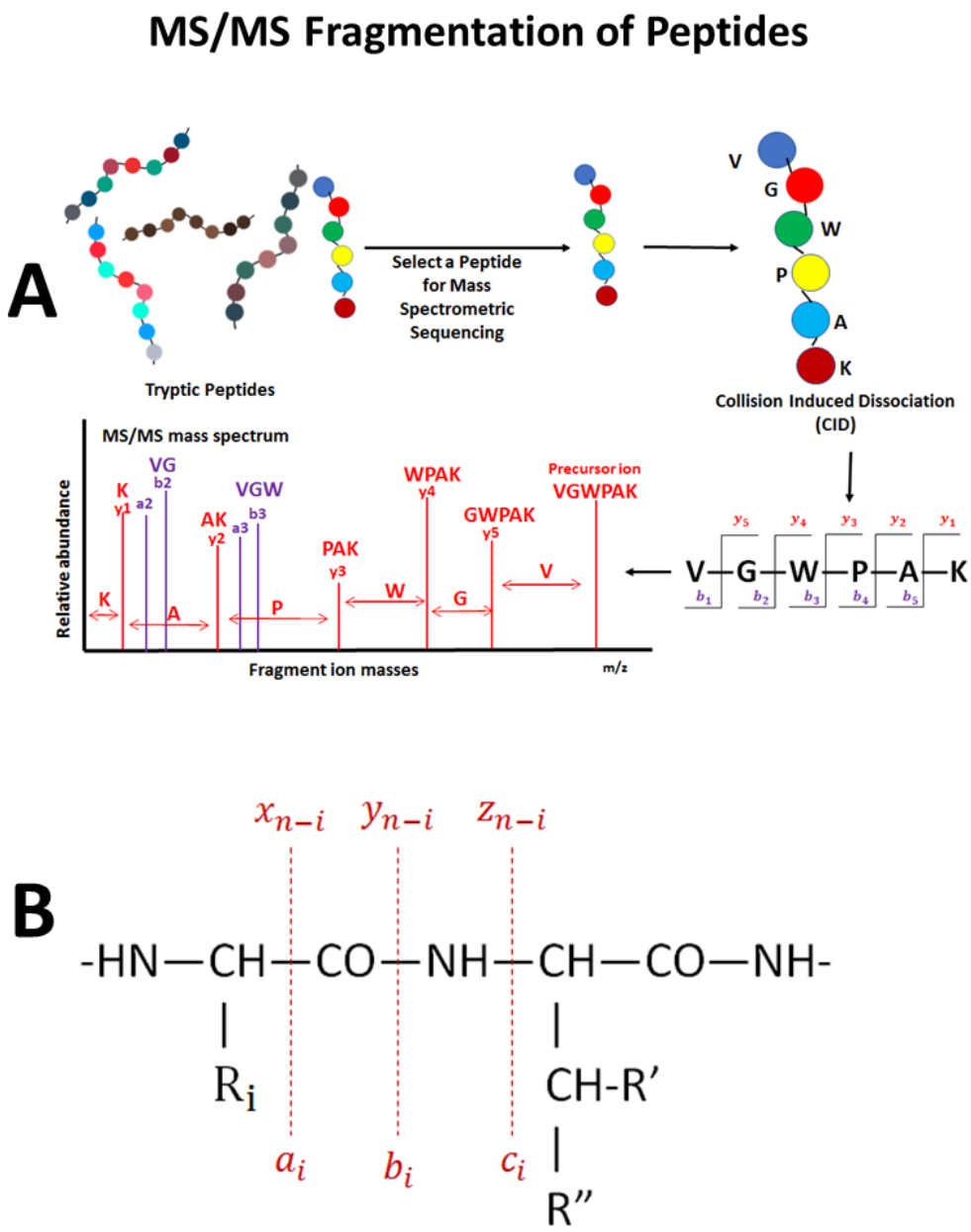


Figure 2. Correspondence between the peptide's amino acid sequence and the fragment ion peaks that are produced in MS/MS (**A**) and the major theoretical fragment ions that are produced from fragmentation of a peptide (**B**).

A particular advantage in proteomics-based analysis is that one protein upon enzymatic digestion can produce many peptides that can be useful in both identification and characterization of that protein. For example, if a protein can theoretically produce 50 peptides that can be analyzed by MS, only a few of those peptides are needed to identify that protein. These peptides can be native peptides and modified peptides (i.e., with Methionine oxidized), or can be native peptides with a different charge state. In addition, identification of these additional charge states for one peptide or unmodified and modified peptides can be particularly useful in characterization of that protein. For example, most therapeutic proteins have at least one or two methionines within their sequences. Knowing whether they are oxidized or not may be useful for long-term storage of those proteins since a methionine does not oxidize if it is buried inside the protein because it is not exposed to solvent. Having two or three charges may be useful in de novo sequencing, when the number of charges indicate the number of amino acids that can be protonated (i.e., only Arg, Lys, and His are/can be protonated).

6.1. The Good

In a typical proteomics experiment, the raw data file is usually dependent on the MS instrument used and it should be first converted into a universal, readable file. More often, this file is a peak list and will concisely show the mass to charge ratio of precursor ions with their relative intensity and charge, as well as the mass to charge ratio of its fragment ions and their relative intensities. OmicX is an online a resource that lists different -omics software. Once the readable file is obtained, it can be submitted for a search against a target database containing theoretical peptide sequences. These sequences are generated by computationally performing theoretical digestions and MS/MS analysis of all the possible proteins/resulting peptides from a given genomic dataset. The search output is a list of peptide-spectrum matches (PSMs) that are used to identify individual peptides. The PSMs are then further used to infer the identity of proteins present in the sample [141]. Such searches can be performed through PLGS or through other programs such as Mascot (Matrix Science, London, UK), which has a subset of programs like the free Mascot Server, Mascot Daemon, Mascot Distiller, and Mascot Parser. This program can be used for protein identification and scoring. In this software, it is also possible to set the parameters to search for certain PTMs based on the mass shift. It is also possible to customize PTM searches to match specific experiments. In 1993, the Yates lab proposed a method that could be used to correlate mass spectral data with a predicted amino acid sequences in a protein database [142]. Following, a program known as Sequest came into the works [143]. PEAKS (Bioinformatics Solutions Inc, Waterloo, ON) is another protein identification software that also has a de novo sequencing function. Comet [144], is another search program which became publicly available in 2012. For experiments requiring relative quantification and statistical analysis, programs such as Scaffold (Proteome Software Inc., Portland, OR, USA) can be utilized. MaxQuant is an important quantitative software that is freely available [145].

Although typical database and spectral library searching are the main tools used for proteomic data analysis, in some instances, de novo sequencing of the raw MS/MS spectrum data is used to identify peptides in instances of novel proteins, mutations, and PTMs. This approach involves acquiring the data and determining mass and composition of peptides directly from the MS/MS spectrum and predicted fragmentation [146]. Current de novo sequencing algorithms include PEAKS [147], Lutefisk [148], PepNovo [149]. Mascot Distiller [150], Protein Prospector [151], Novor [152], UniNovo [153], and PeptideProphet [154].

Proteogenomics, first seen in 2004, in an approach that uses genomic and transcriptomic sequence information as a reference for MS/MS spectra to identify novel peptides [155]. Some examples of proteogenomic approaches include six-frame translation, ab initio gene prediction, and expressed sequence tags. The ability to pair genomic, transcriptomic, and proteomic methods has been the key factor in many discoveries [156–159].

The alignment of amino acids in proteins from different species can be a beneficial characteristic in proteomic analysis. If a well-developed protein database does not exist for a certain species, protein

databases of similar species can be used. The protein hits that are obtained from these other databases can be used as a reference to determine what proteins are present in the species of interest; therefore, databases can be utilized that may not be specifically for the species of interest [160,161]. PEAKS software has a SPIDER algorithm that allows for cross-species homology search and detection of peptide mutations [147].

Many different approaches exist for quantitative proteomics experiments. In a simple, label-free method, a known concentration of an internal standard peptide can be spiked into samples prior to mass spectrometry analysis [148]. Likewise, an external standard peptide sample can be run in between samples on the mass spectrometer, and if the spectral count intensities of the standard runs are the same, the spectral counts/intensities of peptides in the actual samples can be compared as a relative quantitation [162–165]. In this case, the standard acts as proof that running conditions are identical enough that relative quantitation using spectral intensities can be used. There are also many available software for label-free quantitative proteomics, including MapQuant, MZmine, MsInspect, OpenMS, MSight, SuperHirn, and MaxLFQ [166–172]. The advantage of these types of analysis is that they are affordable and time-efficient. This type of quantitation works well enough for pilot and preliminary studies, but findings should be confirmed later using a more precise, reliable method. Labeling methods for quantification are also abundant, including methods such as isotope-coded affinity tag (ICAT), stable isotope labeling by amino acids in cell culture (SILAC), $^{15}N/^{14}N$ metabolic labeling, $^{18}O/^{16}O$ enzymatic labeling, isotope coded protein labeling (ICPL), tandem mass tags (TMT), and isobaric tags for relative and absolute quantification (iTRAQ) [173]. The absolute quantification (AQUA) method allows for the precise determination of protein expression and even post-translational modification levels by mimicking the exact peptide of interest, with the exception of stable isotope enrichment [174].

6.2. The Bad

The aforementioned type of proteomic data analysis is based off the premise that peptides are identified by matching their m/z information to a library of known proteins and their theoretical fragmentation. Similarly, peptides can be identified by matching MS/MS spectra against a library of theoretical spectra for that cell, tissue, organism, etc. which is a slowly growing practice. Therefore, it is assumed that all protein-coding gene sequences are known and well-annotated, and that all proteins encoded by these sequences exist in a database. Such databases include NCBI RefSeq or UniProtKB [175]. Unfortunately, this assumption is far from reality. The PTMs are not easily identifiable from genetic data; therefore, most modified peptides for the proteome of most organisms do not exist in a protein database, which makes proteomic data analysis a tricky task to tackle. As previously stated, de novo sequencing is a method that can be used in this instance. Although de novo sequencing is a very thorough type of analysis, problems with this technique include incomplete peptide fragmentation and low mass accuracy. These problems have been improved slightly in the most current software, but not totally. The downfall of proteogenomic approaches like six-frame translation, ab initio gene prediction, and expressed sequence tags is that they make very large reference databases that can sometimes be hard to work with [175].

Although the alignment of amino acids between different species can be beneficial, it can also be a complication in data analysis because most peptides are not unique to one specific protein or proteome. The NCBI Basic Local Alignment Search Tool (BLAST) shows regions of similarity between protein sequences [176]. Proteins of two different species can have more than 99% alignment of amino acids if they are closely related and can still have much more than 80% alignment if they are distantly related. The quality of alignment and conservation of sequences between proteins is interesting because it can hint towards species homology; however, it can be a complication when analyzing proteomics data.

When trying to perform quantitative proteomics experiments, the tools needed for a reliable analysis can be expensive. The cheapest way to quantify peptides is a label-free approach; however, variable running conditions for the mass spectrometer can lead to errors [177]. The amount of

instrument time required for technical replicates is also a downfall when it comes to label-free approaches. In addition, it is not possible to accurately quantify the relative abundance of different peptides since these molecules have different physicochemical properties and thus will behave differently under mass spectrometry conditions [178]. The different labeling methods for quantitative proteomics also have their flaws. For example, ^{15}N labeling relies on the number of nitrogen atoms present in the proteins, which is variable and can complicate analysis. A disadvantage of the SILAC method is that its cell culture origin limits its capabilities [178]. A universal problem with most labeling approaches is that there may be incomplete labeling of the peptides, which would cause errors in the quantitative analysis.

6.3. The Future

One of the major problems that needs to be addressed in proteomic data analysis is the lack of well-annotated protein databases. It is necessary for researchers to help develop these databases so that they can be used in future proteomics studies. When using a reference database for proteomic analyses, strict, relevant search parameters should be followed to decrease the likelihood of false positives. Whenever possible, smaller protein databases should be used instead of larger ones to minimize the false positives, peptide thresholds should be set to a reasonable tolerance, and relevant PTMs should be taken into consideration. In addition, if peptide identifications are particularly important, it is wise to verify them using the raw data spectrums.

7. False Positives, False Negatives, and Unassigned Spectra

The final and most biologically relevant result of a proteomics experiment is the identification of the proteins present in the sample. This is a two-step process that includes finding PSMs and inferring protein identity based on sequence alignments. Each step is dependent on a database, and, in each step, as the database grows larger, the probability to find peptides and hence proteins due to chance alone increases. Hits that are not actually present in the sample are called false positive identifications and used to be the most challenging problem in proteomics.

7.1. The Good

In order to assess the quality of the PSMs and hence peptide identity, the false discovery rate (FDR) was introduced as the ratio between false PSMs and total number of PSMs. The lower the FDR, the more meaningful the results. Two major strategies for calculating FDR are available: target-decoy search strategy (TDS) [179] and mixture model-based methods [180].

In TDS, fictious decoy peptides are fabricated by reversing or shuffling the protein sequence from the target database. This decoy database is also used for searches, and the FDR level is controlled based on the amount of decoy peptides found in the search output [181]. This approach is the most widely used, although it has several disadvantages. First, as the target database grows larger, so does the decoy database, which increases search time and complexity. Some strategies for reducing [182] or eliminating the decoy database altogether [183] are available and have yet to find their way into the mainstream. Secondly, decoy database searches are more often biased by different peptide features or score functions [181], but some work has been done to overcome this issue [182].

When it comes to protein inference, many effective protein inference algorithms have been developed, such as ProteinProphet, ComByne, and MSBayesPro. For quality assessment of protein identification, two methodologies are available: *p*-value-based approaches that calculate a probability parameter for each identified protein, and FDR approaches that apply a single threshold to all proteins identified [141]. Previously, one of the major sources of false positive protein identification was identification of peptides based on a reduced number of fragment ions. However, this is no longer a problem, since identification of a protein based on one peptide alone is no longer acceptable within the scientific community. This strategy clearly eliminates true identification of some proteins, but it also eliminates any doubts in the proteomics dataset.

The most commonly used computational strategy is to use database search algorithms based on searching acquired MS/MS spectra against a protein sequence database [184]. However, a large fraction of spectra remains unexplained despite a substantial improvement in the quality of MS/MS data acquired on modern mass spectrometers [184]. Novel sequences and the vast diversity of PTMSs remain unidentified in traditional database searches [185]. To develop a fast computational strategy that could be broadly applicable for searches using wide precursor mass tolerance of hundreds of Daltons (open database search), a novel fragment-ion indexing method was designed that provides orders of magnitude improvement in speed over existing tools [186]. This method implements a new database search tool called MS Fragger, which can perform open searches with variable modifications, thus identifying peptides with unknown alterations. It further allows open searches for data sets containing millions of MS/MS spectra and is applicable to data obtained from labeling-based qualitative proteomics experiments [186,187]. Open database searching offers a potential solution to the problem of inaccurate FDR estimates due to unaccounted peptide modifications in traditional narrow mass-window searches [186]. Using an independent platform tool like MS Fragger that is not limited to data from a particular MS instrument can be easily incorporated into most of the existing data analysis pipelines [186].

7.2. The Bad

A problem that will always persist in the proteomics community is false-negative identification of peptides, i.e., not detecting a peptide and a protein that contains it, for a variety of reasons. One such reason is physiological/pathological. For example, if a peptide should be identified in a sample, but it is physiologically/pathologically post-translationally modified (PTMs, i.e., phosphorylation, acetylation etc.), then the peptide is not found in any database search. While most researchers do search for common PTMs in peptides such as phosphorylation, they do not necessarily search for rare PTMs such as farnesylation, and are unable to search for unknown PTMs.

While this problem could be solved (i.e., by expanding the database search using even more powerful software), the price is relatively high and each search will require a considerable amount of time. Another factor for false-negative identification of peptides is experimental. For example, if a peptide that should be identified is modified during the sample preparation process by artificial modifications/PTMs such as methionine oxidation, tryptophan oxidation, cysteine oxidation (to cysteic acid), iodoacetamide-based modification of cysteine (to carbamydomethyl cysteine) or cysteineless peptides, or acrylamide-based modification of cysteine-containing peptides (to propionamide), then these peptides will no longer be identified in a database search, unless all these artificial PTMs are selected in a database search.

Physicochemical properties of proteins and peptides are also important and are not always considered. Hydrophobic proteins and natural PTMs in proteins are two examples of this. For example, hydrophobic peptides do not ionize well and, therefore, are not even found in the MS data. PTMs such as glycosylation and disulfide bridges can prevent trypsin cleavage and thus lead to false negative identification of peptides. An example of a peptide with a series of disulfide bridges, in which analysis of the precursor ions works well, but the disulfide bridges prevent a good fragmentation of these precursor ions, and thus leads to poor quality MS/MS spectra, was recently published from our lab [188].

Incomplete enzymatic digestions (usually trypsin) can lead to erroneous proteomics results and false negative identifications of some peptides. One such case is when one or more trypsin cleavage sites are missed. While 1–3 missed cleavages can easily be fixed (i.e., by database search with missed cleavages), peptides larger than 4–5 kDa that are produced by incomplete trypsin digestion are usually not identified, unless a specialized database is used. Furthermore, fragmentation of such large precursor ions does not produce good MS/MS spectra and therefore does not lead to the positive identification of any proteins [132].

Perhaps one of the biggest problems in a proteomics experiment is the unassigned MS/MS spectra (unpublished observations and [132]). Indeed, in a proteomics experiment, about 60–70% of the spectra are not assigned with confidence to any peptide (unpublished observations). This is true for experiments analyzing just one protein as well as analysis of a large set of proteins (unpublished observations). Again, multiple database searches that are sometimes customized can reduce the number of unassigned peptides to perhaps 50% or as low as 40% in the best-case scenarios, but there is still a lot of work that needs to be done in the area of bioinformatics and database searches.

7.3. The Future

False positive results, false negative results, and unassigned spectra can be solved only by advances in database search and superior bioinformatics analysis of proteomes, i.e., additional software that allows identification of additional peptides that correspond to previously unassigned MS/MS spectra.

8. Conclusions

MS and proteomic methods have greatly advanced the field of protein analysis for both fundamental and clinical research. There are well-established methods for sample preparation and great instrumentation (both UPLC and MS) for MS analysis. There is also a diverse variety of database search engines. However, under the assumption that a proteomics experiment is done properly, the greatest challenge that we still face is a full, comprehensive database search of a proteomics dataset. It is clear that what proteomics needs most is a better, more powerful bioinformatics support for database searching.

Funding: This research received no external funding.

Acknowledgments: We thank the past and current members of the Biochemistry & Proteomics Group for the pleasant working environment.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schwenk, J.M.; Omenn, G.S.; Sun, Z.; Campbell, D.S.; Baker, M.S.; Overall, C.M.; Aebersold, R.; Moritz, R.L.; Deutsch, E.W. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res.* **2017**, *16*, 4299–4310. [CrossRef] [PubMed]
2. Chen, D.; Li, X.Y.; Zhao, X.; Qin, Y.S.; Zhang, X.X.; Li, J.; Wang, J.M.; Wang, C.F. Proteomics and microstructure profiling of goat milk protein after homogenization. *J. Dairy Sci.* **2019**, *102*, 3839–3850. [CrossRef] [PubMed]
3. Mohanta, T.K.; Khan, A.; Hashem, A.; Abd Allah, E.F.; Al-Harrasi, A. The molecular mass and isoelectric point of plant proteomes. *BMC Genom.* **2019**, *20*, 631. [CrossRef] [PubMed]
4. Aslam, B.; Basit, M.; Nisar, M.A.; Khurshid, M.; Rasool, M.H. Proteomics: Technologies and Their Applications. *J. Chromatogr. Sci.* **2017**, *55*, 182–196. [CrossRef]
5. Swinbanks, D. Australia Backs Innovation, Shuns Telescope. *Nature* **1995**, *378*, 653. [CrossRef]
6. Corthals, G.L.; Wasinger, V.C.; Hochstrasser, D.F.; Sanchez, J.C. The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis* **2000**, *21*, 1104–1115. [CrossRef]
7. Anderson, N.G.; Matheson, A.; Anderson, N.L. Back to the future: The human protein index (HPI) and the agenda for post-proteomic biology. *Proteomics* **2001**, *1*, 3–12. [CrossRef]
8. Sperling, K. From proteomics to genomics. *Electrophoresis* **2001**, *22*, 2835–2837. [CrossRef]
9. Loo, J.A.; DeJohn, D.E.; Du, P.; Stevenson, T.I.; Loo, R.R.O. Application of mass spectrometry for target identification and characterization. *Med. Res. Rev.* **1999**, *19*, 307–319. [CrossRef]
10. Aicher, L.; Wahl, D.; Arce, A.; Grenet, O.; Steiner, S. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* **1998**, *19*, 1998–2003. [CrossRef]
11. Zhang, N.; Aebersold, R.; Schwilkowski, B. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2*, 1406–1412. [CrossRef]

12. Tsai, P.L.; Chen, S.F. A Brief Review of Bioinformatics Tools for Glycosylation Analysis by Mass Spectrometry. *Mass Spectrom* **2017**, *6*, S0064. [CrossRef] [PubMed]
13. Nefedov, A.V.; Gilski, M.J.; Sadygov, R.G. Bioinformatics Tools for Mass Spectrometry-Based High-Throughput Quantitative Proteomics Platforms. *Curr. Proteom.* **2011**, *8*, 125–137. [CrossRef]
14. Woods, A.G.; Sokolowska, I.; Ngounou Wetie, A.G.; Channaveerappa, D.; Dupree, E.J.; Jayathirtha, M.; Aslebagh, R.; Wormwood, K.L.; Darie, C.C. Mass Spectrometry for Proteomics-Based Investigation. *Adv. Exp. Med. Biol.* **2019**, *1140*, 1–26. [PubMed]
15. Schirle, M.; Bantscheff, M.; Kuster, B. Mass spectrometry-based proteomics in preclinical drug discovery. *Chem. Biol.* **2012**, *19*, 72–84. [CrossRef]
16. Crutchfield, C.A.; Thomas, S.N.; Sokoll, L.J.; Chan, D.W. Advances in mass spectrometry-based clinical biomarker discovery. *Clin. Proteom.* **2016**, *13*, 1. [CrossRef]
17. Tong, M.; Liu, H.; Hao, J.; Fan, D. Comparative pharmacoproteomics reveals potential targets for berberine, a promising therapy for colorectal cancer. *Biochem. Biophys. Res. Commun.* **2020**, *525*, 244–250. [CrossRef]
18. Parsons, J.; Francavilla, C. 'Omics Approaches to Explore the Breast Cancer Landscape. *Front Cell Dev. Biol.* **2019**, *7*, 395. [CrossRef]
19. Saleem, S.; Tariq, S.; Aleem, I.; Sadr-Ul, S.; Tahseen, M.; Atiq, A.; Hassan, S.; Abu Bakar, M.; Khattak, S.; Syed, A.A.; et al. Proteomics analysis of colon cancer progression. *Clin. Proteom.* **2019**, *16*, 44. [CrossRef]
20. Rogers, J.C.; Bomgarden, R.D. Sample preparation for mass spectrometry-based proteomics; from proteomes to peptides. In *Advances in Experimental Medicine and Biology*; Springer New York LLC: New York, NY, USA, 2016; Volume 919, pp. 43–62.
21. Cristobal, A.; Marino, F.; Post, H.; Van Den Toorn, H.W.P.; Mohammed, S.; Heck, A.J.R. Toward an Optimized Workflow for Middle-Down Proteomics. *Anal. Chem.* **2017**, *89*, 3318–3325. [CrossRef]
22. Kim, M.S.; Pinto, S.M.; Getnet, D.; Nirujogi, R.S.; Manda, S.S.; Chaerkady, R.; Madugundu, A.K.; Kelkar, D.S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581. [CrossRef] [PubMed]
23. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A.M.; Lieberenz, M.; Savitski, M.M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582–587. [CrossRef] [PubMed]
24. Addona, T.A.; Abbatiello, S.E.; Schilling, B.; Skates, S.J.; Mani, D.R.; Bunk, D.M.; Spiegelman, C.H.; Zimmerman, L.J.; Ham, A.J.L.; Keshishian, H.; et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **2009**, *27*, 633–641. [CrossRef] [PubMed]
25. Gillet, L.C.; Leitner, A.; Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem.* **2016**, *9*, 449–472. [CrossRef]
26. Fenyö, D.; Beavis, R.C. The GPMDB REST interface. *Bioinformatics* **2015**, *31*, 2056–2058. [CrossRef]
27. Kiser, J.Z.; Post, M.; Wang, B.; Miyagi, M. Streptomyces erythraeus trypsin for proteomics applications. *J. Proteome Res.* **2009**, *8*, 1810–1817. [CrossRef]
28. Tsiatsiani, L.; Heck, A.J.R. Proteomics beyond trypsin. *FEBS J.* **2015**, *282*, 2612–2626. [CrossRef]
29. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207. [CrossRef]
30. Graves, P.R.; Haystead, T.A.J. Molecular Biologist's Guide to Proteomics. *Microbiol. Mol. Biol. Rev.* **2002**, *66*, 39–63. [CrossRef]
31. Lovrić, J. *Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis*; Wiley-Blackwell: Hoboken, NJ, USA, 2011; p. 283.
32. Futcher, B.; Latter, G.I.; Monardo, P.; McLaughlin, C.S.; Garrels, J.I. A Sampling of the Yeast Proteome. *Mol. Cell. Biol.* **1999**, *19*, 7357–7368. [CrossRef]
33. Kim, H.; Dan, K.; Shin, H.; Lee, J.; Wang, J.I.; Han, D. An efficient method for high-pH peptide fractionation based on C18 StageTips for in-depth proteome profiling. *Anal. Methods* **2019**, *11*, 4693–4698. [CrossRef]
34. Feist, P.; Hummon, A.B. Proteomic challenges: Sample preparation techniques for Microgram-Quantity protein analysis from biological samples. *Int. J. Mol. Sci.* **2015**, *16*, 3537–3563. [CrossRef]
35. Bandow, J.E. Comparison of protein enrichment strategies for proteome analysis of plasma. *Proteomics* **2010**, *10*, 1416–1425. [CrossRef] [PubMed]
36. Polaskova, V.; Kapur, A.; Khan, A.; Molloy, M.P.; Baker, M.S. High-abundance protein depletion: Comparison of methods for human plasma biomarker discovery. *Electrophoresis* **2010**, *31*, 471–482. [CrossRef] [PubMed]

37. Deeb, S.J.; Cox, J.; Schmidt-Supprian, M.; Mann, M. N-linked glycosylation enrichment for in-depth cell surface proteomics of diffuse large B-cell lymphoma subtypes. *Mol. Cell. Proteom.* **2014**, *13*, 240–251. [CrossRef] [PubMed]
38. Nilsson, C.L.; Dillon, R.; Devakumar, A.; Shi, S.D.H.; Greig, M.; Rogers, J.C.; Krastins, B.; Rosenblatt, M.; Kilmer, G.; Major, M.; et al. Quantitative phosphoproteomic analysis of the STAT3/IL-6/HIF1α signaling network: An initial study in GSC11 glioblastoma stem cells. *J. Proteome Res.* **2010**, *9*, 430–443. [CrossRef] [PubMed]
39. Adachi, J.; Kishida, M.; Watanabe, S.; Hashimoto, Y.; Fukamizu, K.; Tomonaga, T. Proteome-wide discovery of unknown ATP-binding proteins and kinase inhibitor target proteins using an ATP probe. *J. Proteome Res.* **2014**, *13*, 5461–5470. [CrossRef]
40. Lemeer, S.; Zörgiebel, C.; Ruprecht, B.; Kohl, K.; Kuster, B. Comparing immobilized kinase inhibitors and covalent ATP probes for proteomic profiling of kinase expression and drug selectivity. *J. Proteome Res.* **2013**, *12*, 1723–1731. [CrossRef]
41. Patricelli, M.P.; Szardenings, A.K.; Liyanage, M.; Nomanbhoy, T.K.; Wu, M.; Weissig, H.; Aban, A.; Chun, D.; Tanner, S.; Kozarich, J.W. Functional interrogation of the kinome using nucleotide acyl phosphates. *Biochemistry* **2007**, *46*, 350–358. [CrossRef]
42. Kim, B.; Araujo, R.; Howard, M.; Magni, R.; Liotta, L.A.; Luchini, A. Affinity enrichment for mass spectrometry: Improving the yield of low abundance biomarkers. *Expert Rev. Proteom.* **2018**, *15*, 353–366. [CrossRef]
43. Ten Have, S.; Boulon, S.; Ahmad, Y.; Lamond, A.I. Mass spectrometry-based immuno-precipitation proteomics—The user's guide. *Proteomics* **2011**, *11*, 1153–1159. [CrossRef] [PubMed]
44. Gundry, R.L.; White, M.Y.; Murray, C.I.; Kane, L.A.; Fu, Q.; Stanley, B.A.; Van Eyk, J.E. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Curr. Protoc. Mol. Biol.* **2009**, *90*, 10–25.
45. Tubaon, R.M.; Haddad, P.R.; Quirino, J.P. Sample Clean-up Strategies for ESI Mass Spectrometry Applications in Bottom-up Proteomics: Trends from 2012 to 2016. *Proteomics* **2017**, *17*, 1700011. [CrossRef] [PubMed]
46. Curreem, S.O.T.; Watt, R.M.; Lau, S.K.P.; Woo, P.C.Y. Two-dimensional gel electrophoresis in bacterial proteomics. *Protein Cell* **2012**, *3*, 346–363. [CrossRef] [PubMed]
47. Kendrick, N.; Darie, C.C.; Hoelter, M.; Powers, G.; Johansen, J. 2D SDS PAGE in Combination with Western Blotting and Mass Spectrometry Is a Robust Method for Protein Analysis with Many Applications. *Adv. Exp. Med. Biol.* **2019**, *1140*, 563–574.
48. Anguraj Vadivel, A.K. Gel-based proteomics in plants: Time to move on from the tradition. *Front. Plant Sci.* **2015**, *6*, 1–4. [CrossRef]
49. Rogowska-Wrzesinska, A.; Le Bihan, M.C.; Thaysen-Andersen, M.; Roepstorff, P. 2D gels still have a niche in proteomics. *J. Proteom.* **2013**, *88*, 4–13. [CrossRef]
50. Loo, R.R.O.; Dales, N.; Andrews, P.C. Surfactant effects on protein structure examined by electrospray ionization mass spectrometry. *Protein Sci.* **1994**, *3*, 1975–1983. [CrossRef]
51. Loo, R.R.; Dales, N.; Andrews, P.C. The effect of detergents on proteins analyzed by electrospray ionization. *Methods Mol. Biol. (Cliftonn. J.)* **1996**, *61*, 141–160.
52. Keller, B.O.; Sui, J.; Young, A.B.; Whittal, R.M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **2008**, *627*, 71–81. [CrossRef]
53. Yocum, A.K.; Yu, K.; Oe, T.; Blair, I.A. Effect of immunoaffinity depletion of human serum during proteomic investigations. *J. Proteome Res.* **2005**, *4*, 1722–1731. [CrossRef] [PubMed]
54. Xu, B.; Zhang, Y.; Zhao, Z.; Yoshida, Y.; Magdeldin, S.; Fujinaka, H.; Ismail, T.A.; Yaoita, E.; Yamamoto, T. Usage of electrostatic eliminator reduces human keratin contamination significantly in gel-based proteomics analysis. *J. Proteom.* **2011**, *74*, 1022–1029. [CrossRef] [PubMed]
55. Hodge, K.; Have, S.T.; Hutton, L.; Lamond, A.I. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteom.* **2013**, *88*, 92–103. [CrossRef] [PubMed]
56. Lyngholm, M.; Vorum, H.; Nielsen, K.; Ehlers, N.; Honoré, B. Attempting to distinguish between endogenous and contaminating cytokeratins in a corneal proteomic study. *BMC Ophthalmol.* **2011**, *11*, 3. [CrossRef]
57. Rabilloud, T. Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains. *Proteomics* **2002**, *2*, 3–10. [CrossRef]

58. Wilkins, M.R.; Gasteiger, E.; Sanchez, J.C.; Bairoch, A.; Hochstrasser, D.F. Two-dimensional gel electrophoresis for proteome projects: The effects of protein hydrophobicity and copy number. *Electrophoresis* **1998**, *19*, 1501–1505. [CrossRef] [PubMed]
59. Doellinger, J.; Schneider, A.; Hoeller, M.; Lasch, P. Sample Preparation by Easy Extraction and Digestion (SPEED)—A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol. Cell. Proteom.* **2020**, *19*, 209–222. [CrossRef] [PubMed]
60. Ye, X.T.; Tang, J.; Mao, Y.H.; Lu, X.; Yang, Y.; Chen, W.D.; Zhang, X.Y.; Xu, R.L.; Tian, R.J. Integrated proteomics sample preparation and fractionation: Method development and applications. *Trac-Trends Anal. Chem.* **2019**, *120*, 115667. [CrossRef]
61. Wiśniewski, J.R. Filter-Aided Sample Preparation: The Versatile and Efficient Method for Proteomic Analysis. In *Methods in Enzymology*; Academic Press Inc.: Cambridge, MA, USA, 2017; pp. 15–27.
62. Wiśniewski, J.R. Filter-aided sample preparation for proteome analysis. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2018; pp. 3–10.
63. Kulak, N.A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **2014**, *11*, 319. [CrossRef]
64. Zhu, Y.; Piehowski, P.D.; Zhao, R.; Chen, J.; Shen, Y.F.; Moore, R.J.; Shukla, A.K.; Petyuk, V.A.; Campbell-Thompson, M.; Mathews, C.E.; et al. Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **2018**, *9*, 1–10. [CrossRef]
65. Yuan, H.M.; Zhang, S.; Zhao, B.F.; Weng, Y.J.; Zhu, X.D.; Li, S.W.; Zhang, L.H.; Zhang, Y.K. Enzymatic Reactor with Trypsin Immobilized on Graphene Oxide Modified Polymer Microspheres To Achieve Automated Proteome Quantification. *Anal. Chem.* **2017**, *89*, 6324–6329. [CrossRef] [PubMed]
66. Zougman, A.; Selby, P.J.; Banks, R.E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **2014**, *14*, 1006–1010. [CrossRef]
67. Hughes, C.S.; Foehr, S.; Garfield, D.A.; Furlong, E.E.; Steinmetz, L.M.; Krijgsveld, J. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **2014**, *10*, 757. [CrossRef]
68. Muller, T.; Kalxdorf, M.; Longuespee, R.; Kazdal, D.N.; Stenzinger, A.; Krijgsveld, J. Automated sample preparation with SP3 for low-input clinical proteomics. *Mol. Syst. Biol.* **2020**, *16*, e9111. [CrossRef] [PubMed]
69. Wolters, D.A.; Washburn, M.P.; Yates, J.R., 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001**, *73*, 5683–5690. [CrossRef] [PubMed]
70. Horning, E.C.; Carroll, D.I.; Dzidic, I.; Haegele, K.D.; Horning, M.G.; Stillwell, R.N. Atmospheric pressure ionization (API) mass spectrometry. Solvent-mediated ionization of samples introduced in solution and in a liquid chromatograph effluent stream. *J. Chromatogr. Sci.* **1974**, *12*, 725–729. [CrossRef]
71. Blakley, C.R.; Carmody, J.J.; Vestal, M.L. A New Soft Ionization Technique for Mass-Spectrometry of Complex-Molecules. *J. Am. Chem. Soc.* **1980**, *102*, 5931–5933. [CrossRef]
72. Blakley, C.R.; Carmody, J.C.; Vestal, M.L. Combined Liquid Chromatograph-Mass Spectrometer for Involatile Biological Samples. *Clin. Chem.* **1980**, *26*, 1467–1473. [CrossRef]
73. Anderson, N.L.; Anderson, N.G. The human plasma proteome—History, character, and diagnostic prospects. *Mol. Cell. Proteom.* **2002**, *1*, 845–867. [CrossRef]
74. Mant, C.T.; Chen, Y.; Yan, Z.; Popa, T.V.; Kovacs, J.M.; Mills, J.B.; Tripet, B.P.; Hodges, R.S. HPLC analysis and purification of peptides. *Methods Mol. Biol.* **2007**, *386*, 3–55.
75. Neverova, I.; Van Eyk, J.E. Application of reversed phase high performance liquid chromatography for subproteomic analysis of cardiac muscle. *Proteomics* **2002**, *2*, 22–31. [CrossRef]
76. Chait, B.T. Chemistry. Mass spectrometry: Bottom-up or top-down? *Science* **2006**, *314*, 65–66. [CrossRef] [PubMed]
77. Duong-Ly, K.C.; Gabelli, S.B. Gel filtration chromatography (size exclusion chromatography) of proteins. *Methods Enzym.* **2014**, *541*, 105–114.
78. Li, Y.; Gu, C.; Gruenhagen, J.; Zhang, K.; Yehl, P.; Chetwyn, N.P.; Medley, C.D. A size exclusion-reversed phase two dimensional-liquid chromatography methodology for stability and small molecule related species in antibody drug conjugates. *J. Chromatogr. A* **2015**, *1393*, 81–88. [CrossRef]
79. Duong-Ly, K.C.; Gabelli, S.B. Using ion exchange chromatography to purify a recombinantly expressed protein. *Methods Enzym.* **2014**, *541*, 95–103.

80. Cai, W.; Tucholski, T.; Chen, B.; Alpert, A.J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y. Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal. Chem.* **2017**, *89*, 5467–5475. [CrossRef]
81. Freeman, W.M.; Lull, M.E.; Guilford, M.T.; Vrana, K.E. Depletion of abundant proteins from non-human primate serum for biomarker studies. *Proteomics* **2006**, *6*, 3109–3113. [CrossRef]
82. Kline, K.G.; Wu, C.C. MudPIT analysis: Application to human heart tissue. *Methods Mol. Biol.* **2009**, *528*, 281–293.
83. Motoyama, A.; Venable, J.D.; Ruse, C.I.; Yates, J.R., 3rd. Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples. *Anal. Chem.* **2006**, *78*, 5109–5118. [CrossRef]
84. Issaq, H.J.; Chan, K.C.; Janini, G.M.; Conrads, T.P.; Veenstra, T.D. Multidimensional separation of peptides for effective proteomic analysis. *.J Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2005**, *817*, 35–47. [CrossRef]
85. Washburn, M.P.; Wolters, D.; Yates, J.R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19*, 242–247. [CrossRef] [PubMed]
86. Lambert, J.P.; Ethier, M.; Smith, J.C.; Figeys, D. Proteomics: From gel based to gel free. *Anal. Chem.* **2005**, *77*, 3771–3787. [CrossRef] [PubMed]
87. Yu, D.; Wang, Z.; Cupp-Sutton, K.A.; Liu, X.; Wu, S. Deep Intact Proteoform Characterization in Human Cell Lysate Using High-pH and Low-pH Reversed-Phase Liquid Chromatography. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2502–2513. [CrossRef] [PubMed]
88. Aguilar, M.I. Reversed-phase high-performance liquid chromatography. *Methods Mol. Biol.* **2004**, *251*, 9–22. [PubMed]
89. Solovyeva, E.M.; Lobas, A.A.; Kopylov, A.T.; Ilina, I.Y.; Levitsky, L.I.; Moshkovskii, S.A.; Gorshkov, M.V. FractionOptimizer: A method for optimal peptide fractionation in bottom-up proteomics. *Anal. Bioanal. Chem.* **2018**, *410*, 3827–3833. [CrossRef]
90. Nogueira, R.; Lammerhofer, M.; Lindner, W. Alternative high-performance liquid chromatographic peptide separation and purification concept using a new mixed-mode reversed-phase/weak anion-exchange type stationary phase. *J. Chromatogr. A* **2005**, *1089*, 158–169. [CrossRef]
91. Phillips, H.L.; Williamson, J.C.; van Elburg, K.A.; Snijders, A.P.; Wright, P.C.; Dickman, M.J. Shotgun proteome analysis utilising mixed mode (reversed phase-anion exchange chromatography) in conjunction with reversed phase liquid chromatography mass spectrometry analysis. *Proteomics* **2010**, *10*, 2950–2960. [CrossRef]
92. Vasconcelos Soares Maciel, E.; de Toffoli, A.L.; Sobieski, E.; Domingues Nazario, C.E.; Lancas, F.M. Miniaturized liquid chromatography focusing on analytical columns and mass spectrometry: A review. *Anal. Chim. Acta* **2020**, *1103*, 11–31. [CrossRef]
93. Zuvela, P.; Skoczylas, M.; Jay Liu, J.; Ba Czek, T.; Kaliszan, R.; Wong, M.W.; Buszewski, B. Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chem. Rev.* **2019**, *119*, 3674–3729. [CrossRef]
94. Maya, F.; Paull, B. Recent strategies to enhance the performance of polymer monoliths for analytical separations. *J. Sep. Sci.* **2019**, *42*, 1564–1576. [CrossRef]
95. Luo, Q.; Shen, Y.; Hixson, K.K.; Zhao, R.; Yang, F.; Moore, R.J.; Mottaz, H.M.; Smith, R.D. Preparation of 20-microm-i.d. silica-based monolithic columns and their performance for proteomics analyses. *Anal. Chem.* **2005**, *77*, 5028–5035. [CrossRef]
96. Novotny, M.V. Capillary biomolecular separations. *J. Chromatogr. B Biomed. Sci. Appl.* **1997**, *689*, 55–70. [CrossRef]
97. Premstaller, A.; Oberacher, H.; Walcher, W.; Timperio, A.M.; Zolla, L.; Chervet, J.P.; Cavusoglu, N.; van Dorsselaer, A.; Huber, C.G. High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies. *Anal. Chem.* **2001**, *73*, 2390–2396. [CrossRef] [PubMed]
98. He, B.; Tait, N.; Regnier, F. Fabrication of nanocolumns for liquid chromatography. *Anal. Chem.* **1998**, *70*, 3790–3797. [CrossRef]
99. Ion, L.; Petre, B.A. Immuno-Affinity Mass Spectrometry: A Novel Approaches with Biomedical Relevance. *Adv. Exp. Med. Biol.* **2019**, *1140*, 377–388. [PubMed]

100. Yuan, X.; Oleschuk, R.D. Advances in Microchip Liquid Chromatography. *Anal Chem* **2018**, *90*, 283–301. [CrossRef] [PubMed]
101. Andjelkovic, U.; Tufegdzic, S.; Popovic, M. Use of monolithic supports for high-throughput protein and peptide separation in proteomics. *Electrophoresis* **2017**, *38*, 2851–2869. [CrossRef]
102. Thakur, S.S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteom.* **2011**, *10*, M110.003699. [CrossRef]
103. Van Eeckhaut, A.; Lanckmans, K.; Sarre, S.; Smolders, I.; Michotte, Y. Validation of bioanalytical LC-MS/MS assays: Evaluation of matrix effects. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2009**, *877*, 2198–2207. [CrossRef]
104. MacNair, J.E.; Lewis, K.C.; Jorgenson, J.W. Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. *Anal. Chem.* **1997**, *69*, 983–989. [CrossRef]
105. Denoroy, L.; Zimmer, L.; Renaud, B.; Parrot, S. Ultra high performance liquid chromatography as a tool for the discovery and the analysis of biomarkers of diseases: A review. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2013**, *927*, 37–53. [CrossRef] [PubMed]
106. Wilson, I.D.; Nicholson, J.K.; Castro-Perez, J.; Granger, J.H.; Johnson, K.A.; Smith, B.W.; Plumb, R.S. High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *J. Proteome Res.* **2005**, *4*, 591–598. [CrossRef] [PubMed]
107. Neue, U.D.; Kele, M.; Bunner, B.; Kromidas, A.; Dourdeville, T.; Mazzeo, J.R.; Grumbach, E.S.; Serpa, S.; Wheat, T.E.; Hong, P.; et al. Ultra-performance liquid chromatography technology and applications. *Adv. Chromatogr.* **2010**, *48*, 99–143.
108. Vuckovic, D.; Dagley, L.F.; Purcell, A.W.; Emili, A. Membrane proteomics by high performance liquid chromatography–tandem mass spectrometry: Analytical approaches and challenges. *Proteomics* **2013**, *13*, 404–423. [CrossRef] [PubMed]
109. Fairchild, J.N.; Horvath, K.; Guiochon, G. Theoretical advantages and drawbacks of on-line, multidimensional liquid chromatography using multiple columns operated in parallel. *J. Chromatogr. A* **2009**, *1216*, 6210–6217. [CrossRef] [PubMed]
110. Ball, C.H.; Roulhac, P.L. Multidimensional Techniques in Protein Separations for Neuroproteomics. In *Neuroproteomics*; Alzate, O., Ed.; CRC Press: Boca Raton, FL, USA, 2010.
111. Kota, U.; Stolowitz, M.L. Improving Proteome Coverage by Reducing Sample Complexity via Chromatography. *Adv. Exp. Med. Biol.* **2016**, *919*, 83–143.
112. Kocova Vlckova, H.; Pilarova, V.; Svobodova, P.; Plisek, J.; Svec, F.; Novakova, L. Current state of bioanalytical chromatography in clinical analysis. *Analyst* **2018**, *143*, 1305–1325. [CrossRef]
113. Ucles Moreno, A.; Herrera Lopez, S.; Reichert, B.; Lozano Fernandez, A.; Hernando Guil, M.D.; Fernandez-Alba, A.R. Microflow liquid chromatography coupled to mass spectrometry–an approach to significantly increase sensitivity, decrease matrix effects, and reduce organic solvent usage in pesticide residue analysis. *Anal. Chem.* **2015**, *87*, 1018–1025. [CrossRef]
114. Spicer, V.; Krokhin, O.V. Peptide retention time prediction in hydrophilic interaction liquid chromatography. Comparison of separation selectivity between bare silica and bonded stationary phases. *J. Chromatogr. A* **2018**, *1534*, 75–84. [CrossRef]
115. Fenn, J.B.; Mann, M.; Meng, C.K.; Wong, S.F.; Whitehouse, C.M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246*, 64–71. [CrossRef]
116. Strathmann, F.G.; Hoofnagle, A.N. Current and Future Applications of Mass Spectrometry to the Clinical Laboratory. *Am. J. Clin. Pathol.* **2011**, *136*, 609–616. [CrossRef] [PubMed]
117. Chalmers, M.J.; Gaskell, S.J. Advances in mass spectrometry for proteome analysis. *Curr. Opin. Biotechnol.* **2000**, *11*, 384–390. [CrossRef]
118. Bateman, N.W.; Goulding, S.P.; Shulman, N.J.; Gadok, A.K.; Szumlinski, K.K.; MacCoss, M.J.; Wu, C.C. Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Mol. Cell. Proteom.* **2014**, *13*, 329–338. [CrossRef] [PubMed]
119. Neilson, K.A.; Ali, N.A.; Muralidharan, S.; Mirzaei, M.; Mariani, M.; Assadourian, G.; Lee, A.; Van Sluyter, S.C.; Haynes, P.A. Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics* **2011**, *11*, 535–553. [CrossRef]

120. Tabb, D.L.; Vega-Montoto, L.; Rudnick, P.A.; Variyath, A.M.; Ham, A.-J.L.; Bunk, D.M.; Kilpatrick, L.E.; Billheimer, D.D.; Blackman, R.K.; Cardasis, H.L. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2010**, *9*, 761–776. [CrossRef] [PubMed]
121. Doerr, A. DIA mass spectrometry. *Nat. Methods* **2014**, *12*, 35. [CrossRef]
122. Bilbao, A.; Varesio, E.; Luban, J.; Strambio-De-Castillia, C.; Hopfgartner, G.; Müller, M.; Lisacek, F. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **2015**, *15*, 964–980. [CrossRef]
123. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **2008**, *4*, 14. [CrossRef]
124. Maiolica, A.; Jünger, M.A.; Ezkurdia, I.; Aebersold, R. Targeted proteome investigation via selected reaction monitoring mass spectrometry. *J. Proteom.* **2012**, *75*, 3495–3513. [CrossRef]
125. Stone, P.; Glauner, T.; Kuhlmann, F.; Schlabach, T.; Miller, K. *New Dynamic MRM Mode Improves Data Quality and Triple Quad Quantification in Complex Analyses*; Agilent Publication: Santa Clara, CA, USA, 2009.
126. Rauniyar, N. Parallel reaction monitoring: A targeted experiment performed using high resolution and high mass accuracy mass spectrometry. *Int. J. Mol. Sci.* **2015**, *16*, 28566–28581. [CrossRef]
127. Ronsein, G.E.; Pamir, N.; von Haller, P.D.; Kim, D.S.; Oda, M.N.; Jarvik, G.P.; Vaisar, T.; Heinecke, J.W. Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics. *J. Proteom.* **2015**, *113*, 388–399. [CrossRef]
128. Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteom.* **2012**, *11*, 1709–1723. [CrossRef] [PubMed]
129. Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteom.* **2011**, *10*, M111.011015. [CrossRef] [PubMed]
130. Bonner, R.; Hopfgartner, G. SWATH data independent acquisition mass spectrometry for metabolomics. *Trac. Trends Anal. Chem.* **2019**, *120*, 115278. [CrossRef]
131. Lenz, E.; Bright, J.; Knight, R.; Westwood, F.; Davies, D.; Major, H.; Wilson, I. Metabonomics with 1H-NMR spectroscopy and liquid chromatography-mass spectrometry applied to the investigation of metabolic changes caused by gentamicin-induced nephrotoxicity in the rat. *Biomark* **2005**, *10*, 173–187. [CrossRef] [PubMed]
132. Channaveerappa, D.; Ngounou Wetie, A.G.; Darie, C.C. Bottlenecks in Proteomics: An Update. *Adv. Exp. Med. Biol.* **2019**, *1140*, 753–769.
133. Morris, C.B.; Poland, J.C.; May, J.C.; McLean, J.A. Fundamentals of Ion Mobility-Mass Spectrometry for the Analysis of Biomolecules. *Methods Mol. Biol.* **2020**, *2084*, 1–31.
134. Distler, U.; Kuharev, J.; Navarro, P.; Tenzer, S. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. *Nat. Protoc.* **2016**, *11*, 795–812. [CrossRef]
135. Rifai, N.; Gillette, M.A.; Carr, S.A. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nat. Biotechnol.* **2006**, *24*, 971–983. [CrossRef]
136. Han, X.; Aslanian, A.; Yates, J.R., 3rd. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12*, 483–490. [CrossRef]
137. Liu, Y.; Buil, A.; Collins, B.C.; Gillet, L.C.J.; Blum, L.C.; Cheng, L.-Y.; Vitek, O.; Mouritsen, J.; Lachance, G.; Spector, T.D.; et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **2015**, *11*, 786. [CrossRef] [PubMed]
138. Contino, N.C.; Pierson, E.E.; Keifer, D.Z.; Jarrold, M.F. Charge Detection Mass Spectrometry with Resolved Charge States. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 101–108. [CrossRef] [PubMed]
139. Zubarev, R.A. Electron-capture dissociation tandem mass spectrometry. *Curr. Opin. Biotechnol.* **2004**, *15*, 12–16. [CrossRef]
140. Mikesh, L.M.; Ueberheide, B.; Chi, A.; Coon, J.J.; Syka, J.E.; Shabanowitz, J.; Hunt, D.F. The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta.* **2006**, *1764*, 1811–1822. [CrossRef] [PubMed]
141. Wu, G.Y.; Wan, X.; Xu, B.H. A new estimation of protein-level false discovery rate. *BMC Genom.* **2018**, *19*, 567. [CrossRef]
142. Eng, J.K.; McCormack, A.L.; Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989. [CrossRef]

143. Sadygov, R.G. Using SEQUEST with Theoretically Complete Sequence Databases. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1858–1864. [CrossRef]
144. Eng, J.K.; Hoopmann, M.R.; Jahan, T.A.; Egertson, J.D.; Noble, W.S.; MacCoss, M.J. A Deeper Look into Comet-Implementation and Features. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1865–1874. [CrossRef]
145. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319. [CrossRef]
146. Webb-Robertson, B.J.M.; Cannon, W.R. Current trends in computational inference from mass spectrometry-based proteomics. *Brief. Bioinform.* **2007**, *8*, 304–317. [CrossRef]
147. Ma, B.; Zhang, K.Z.; Hendrie, C.; Liang, C.Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337–2342. [CrossRef] [PubMed]
148. Taylor, J.A.; Johnson, R.S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067–1075. [CrossRef]
149. Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77*, 964–973. [CrossRef] [PubMed]
150. Kalb, S.R.; Baudys, J.; Rees, J.C.; Smith, T.J.; Smith, L.A.; Helma, C.H.; Hill, K.; Kull, S.; Kirchner, S.; Dorner, M.B.; et al. De novo subtype and strain identification of botulinum neurotoxin type B through toxin proteomics. *Anal. Bioanal. Chem.* **2012**, *403*, 215–226. [CrossRef]
151. Medzihradszky, K.F.; Chalkley, R.J. Lessons in De Novo Peptide Sequencing by Tandem Mass Spectrometry. *Mass Spectrom. Rev.* **2015**, *34*, 43–63. [CrossRef]
152. Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1885–1894. [CrossRef] [PubMed]
153. Jeong, K.; Kim, S.; Pevzner, P.A. UniNovo: A universal tool for de novo peptide sequencing. *Bioinformatics* **2013**, *29*, 1953–1962. [CrossRef]
154. Keller, A.; Nesvizhskii, A.I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392. [CrossRef] [PubMed]
155. Jaffe, J.D.; Berg, H.C.; Church, G.M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004**, *4*, 59–77. [CrossRef] [PubMed]
156. Fertig, E.J.; Slebos, R.; Chung, C.H. Application of genomic and proteomic technologies in biomarker discovery. *Am. Soc. Clin. Oncol. Educ. Book* **2012**, *1*, 377–382. [CrossRef]
157. Sheynkman, G.M.; Shortreed, M.R.; Cesnik, A.J.; Smith, L.M. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu. Rev. Anal. Chem. (Palo Alto. Calif.)* **2016**, *9*, 521–545. [CrossRef] [PubMed]
158. Lai, Y.J.; Yu, D.; Zhang, J.H.; Chen, G.J. Cooperation of Genomic and Rapid Nongenomic Actions of Estrogens in Synaptic Plasticity. *Mol. Neurobiol.* **2017**, *54*, 4113–4126. [CrossRef] [PubMed]
159. Whiteaker, J.R.; Halusa, G.N.; Hoofnagle, A.N.; Sharma, V.; MacLean, B.; Yan, P.; Wrobel, J.A.; Kennedy, J.; Mani, D.R.; Zimmerman, L.J.; et al. CPTAC Assay Portal: A repository of targeted proteomic assays. *Nat. Methods* **2014**, *11*, 703–704. [CrossRef] [PubMed]
160. Dupree, E.J.; Crimmins, B.S.; Holsen, T.M.; Darie, C.C. Developing Well-Annotated Species-Specific Protein Databases Using Comparative Proteogenomics. *Adv. Exp. Med. Biol.* **2019**, *1140*, 389–400. [PubMed]
161. Dupree, E.J.; Crimmins, B.S.; Holsen, T.M.; Darie, C.C. Proteomic Analysis of the Lake Trout (Salvelinus namaycush) Liver Identifies Proteins from Evolutionarily Close and -Distant Fish Relatives. *Proteomics* **2019**, *19*, e1800429. [CrossRef]
162. Mihasan, M.; Babii, C.; Aslebagh, R.; Channaveerappa, D.; Dupree, E.; Darie, C.C. Proteomics based analysis of the nicotine catabolism in Paenarthrobacter nicotinovorans pAO1. *Sci. Rep.* **2018**, *8*, 16239. [CrossRef]
163. Aslebagh, R.; Channaveerappa, D.; Arcaro, K.F.; Darie, C.C. Proteomics analysis of human breast milk to assess breast cancer risk. *Electrophoresis* **2018**, *39*, 653–665. [CrossRef]
164. Channaveerappa, D.; Lux, J.C.; Wormwood, K.L.; Heintz, T.A.; McLerie, M.; Treat, J.A.; King, H.; Alnasser, D.; Goodrow, R.J.; Ballard, G.; et al. Atrial electrophysiological and molecular remodelling induced by obstructive sleep apnoea. *J. Cell. Mol. Med.* **2017**, *21*, 2223–2235. [CrossRef]
165. Ngounou Wetie, A.G.; Wormwood, K.L.; Russell, S.; Ryan, J.P.; Darie, C.C.; Woods, A.G. A Pilot Proteomic Analysis of Salivary Biomarkers in Autism Spectrum Disorder. *Autism. Res.* **2015**, *8*, 338–350. [CrossRef]

166. Leptos, K.C.; Sarracino, D.A.; Jaffe, J.D.; Krastins, B.; Church, G.M. MapQuant: Open-source software for large-scale protein quantification. *Proteomics* **2006**, *6*, 1770–1782. [CrossRef]
167. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, 395. [CrossRef] [PubMed]
168. Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.H.; Lin, C.W.; et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **2006**, *22*, 1902–1909. [CrossRef] [PubMed]
169. Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. OpenMS-An open-source software framework for mass spectrometry. *BMC Bioinform.* **2008**, *9*, 1–11. [CrossRef] [PubMed]
170. Palagi, P.M.; Walther, D.; Quadroni, M.; Catherinet, S.; Burgess, J.; Zimmermann-Ivol, C.G.; Sanchez, J.C.; Binz, P.A.; Hochstrasser, D.F.; Appel, R.D. MSight: An image analysis software for liquid chromatography-mass spectrometry. *Proteomics* **2005**, *5*, 2381–2384. [CrossRef] [PubMed]
171. Mueller, L.N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.Y.; Vitek, O.; Aebersold, R.; Muller, M. SuperHirn—A novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007**, *7*, 3470–3480. [CrossRef] [PubMed]
172. Cox, J.; Hein, M.Y.; Luber, C.A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteom.* **2014**, *13*, 2513–2526. [CrossRef] [PubMed]
173. Zhu, W.H.; Smith, J.W.; Huang, C.M. Mass Spectrometry-Based Label-Free Quantitative Proteomics. *J. Biomed. Biotechnol.* **2010**, *2010*, 840518. [CrossRef]
174. Kirkpatrick, D.S.; Gerber, S.A.; Gygi, S.P. The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. *Methods* **2005**, *35*, 265–273. [CrossRef]
175. Nesvizhskii, A.I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114–1125. [CrossRef]
176. Matsuda, F.; Tsugawa, H.; Fukusaki, E. Method for Assessing the Statistical Significance of Mass Spectral Similarities Using Basic Local Alignment Search Tool Statistics. *Anal. Chem.* **2013**, *85*, 8291–8297. [CrossRef]
177. Old, W.M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K.G.; Mendoza, A.; Sevinsky, J.R.; Resing, K.A.; Ahn, N.G. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteom.* **2005**, *4*, 1487–1502. [CrossRef] [PubMed]
178. Chen, X.L.; Wei, S.S.; Ji, Y.L.; Guo, X.J.; Yang, F.Q. Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics* **2015**, *15*, 3175–3192. [CrossRef] [PubMed]
179. Elias, J.E.; Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207–214. [CrossRef] [PubMed]
180. Choi, H.; Nesvizhskii, A.I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254–265. [CrossRef]
181. Danilova, Y.; Voronkova, A.; Sulimov, P.; Kertesz-Farkas, A. Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification. *J. Proteome Res.* **2019**, *18*, 2354–2358. [CrossRef]
182. Kim, H.; Lee, S.; Park, H. Target-small decoy search strategy for false discovery rate estimation. *BMC Bioinform.* **2019**, *20*, 438. [CrossRef]
183. Gonnelli, G.; Stock, M.; Verwaeren, J.; Maddelein, D.; De Baets, B.; Martens, L.; Degroeve, S. A Decoy-Free Approach to the Identification of Peptides. *J. Proteome Res.* **2015**, *14*, 1792–1798. [CrossRef]
184. Eng, J.K.; Searle, B.C.; Clauser, K.R.; Tabb, D.L. A face in the crowd: Recognizing peptides through database search. *Mol. Cell. Proteomics* **2011**, *10*. [CrossRef]
185. Skinner, O.S.; Kelleher, N.L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **2015**, *33*, 717. [CrossRef]
186. Kong, A.T.; Leprevost, F.V.; Avtonomov, D.M.; Mellacheruvu, D.; Nesvizhskii, A.I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **2017**, *14*, 513. [CrossRef]

187. Meyer, J.G. Fast Proteome Identification and Quantification from Data-Dependent Acquisition-Tandem Mass Spectrometry using Free Software Tools. *Methods Protoc.* **2019**, *2*, 8. [CrossRef] [PubMed]
188. Wormwood, K.L.; Ngounou Wetie, A.G.; Gomez, M.V.; Ju, Y.; Kowalski, P.; Mihasan, M.; Darie, C.C. Structural Characterization and Disulfide Assignment of Spider Peptide Phalpha1beta by Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 827. [CrossRef] [PubMed]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

Computational and Statistical Analysis of Protein Mass Spectrometry Data

William Stafford Noble[1,2]*, Michael J. MacCoss[1]

1 Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **2** Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Abstract: High-throughput proteomics experiments involving tandem mass spectrometry produce large volumes of complex data that require sophisticated computational analyses. As such, the field offers many challenges for computational biologists. In this article, we briefly introduce some of the core computational and statistical problems in the field and then describe a variety of outstanding problems that readers of *PLoS Computational Biology* might be able to help solve.

This is an "Editors' Outlook" article for *PLoS Computational Biology*

Introduction

DNA gets a lot of attention these days, in whole genome sequencing projects, genome-wide association studies, and experiments measuring transcription factor binding, chromatin accessibility, DNA methylation, and histone modification profiles. But proteins are the molecular workhorses of the cell, and proteomics—the systematic study of the complete set of proteins expressed in a given cell, tissue, or organism—is poised to become the next hot topic.

Just as next-generation sequencing tehcnology is driving the current genomics boom, so improvements in tandem mass spectrometry technology are leading to more comprehensive and precise proteomics assays. Like a short-read sequencing machine, a mass spectrometer runs 24 hours a day, producing a huge quantity of data. And like short-read sequencing data, mass spectral data sets exhibit complex dependencies and patterns of missing data. In both fields, the underlying technologies, along with the characteristics of the resulting data sets, change rapidly, requiring constant development of new analytical methods.

Strikingly, however, relatively few bioinformatics researchers work on methods for analyzing mass spectrometry data. *PLoS Computational Biology* published only two papers on the topic in 2010 [1,2]. At the Intelligent Systems for Molecular Biology (ISMB) conference, none of the designated subject areas for submitted manuscripts is relevant to mass spectrometry analysis, and over the last three years, a total of four mass spectometry papers were published at ISMB, each appearing in the "Other bioinformatics applications" category. Meanwhile, the annual American Society for Mass Spectrometry conference draws more than 6,000 attendees, and the society boasts 7,000 current members.

Some of the forces preventing people from entering the mass spectrometry research arena are social. As a field, mass spectrometry is older than genomics, and as such, the norms around the sharing of mass spectrometry data are less open. In addition, intellectual property issues, such as the SEQUEST patents held by the University of Washington, may have discouraged some researchers from entering the field.

An equally important impediment, however, is the "energy barrier" associated with starting out in mass spectrometry. In the late 1990s, microarray analysis took off with surprising rapidity, in part because the data could be fairly accurately summarized in matrix format, and manipulating matrices is familiar to computer scientistics, electrical engineers, and physicists. This is not the case with mass spectrometry data. If you show someone with no relevant background knowledge a DNA sequence or a microarray image, explaining what they are looking at will take less time than explaining to that same person what a peptide fragmentation spectrum is.

The goal of this article is to lower that energy barrier by explaining in simple terms how a tandem mass spectrometry experiment works and what are the key research problems associated with this type of data.

A Typical Shotgun Proteomics Experiment

A typical shotgun proteomics experiment proceeds in three steps, as illustrated in Figure 1A. The input to the experiment is a collection of proteins, which have been isolated from a complex mixture. A typical complex mixture may contain a few thousand proteins, ranging in abundance from tens of copies to hundreds of thousands of copies.

In the first experimental step, the proteins are digested into peptides using a protease. This digestion is necessary because whole proteins are too massive to be subject to direct mass spectometry analysis without using very expensive equipment. Second, the peptides are subjected to liquid chromatography, in which the peptides pass through a thin glass column that separates the peptides based on a particular chemical property (e.g., the hydrophobicity). This separation step reduces the complexity of the mixtures of peptides going into the mass spectrometer. The

Citation: Noble WS, MacCoss MJ (2012) Computational and Statistical Analysis of Protein Mass Spectrometry Data. PLoS Comput Biol 8(1): e1002296. doi:10.1371/journal.pcbi.1002296

Editor: Philip E. Bourne, University of California San Diego, United States of America

Published January 26, 2012

Copyright: © 2012 Noble, MacCoss. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by National Institutes of Health awards R01 GM096306 and P41 RR0011823. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: william-noble@uw.edu

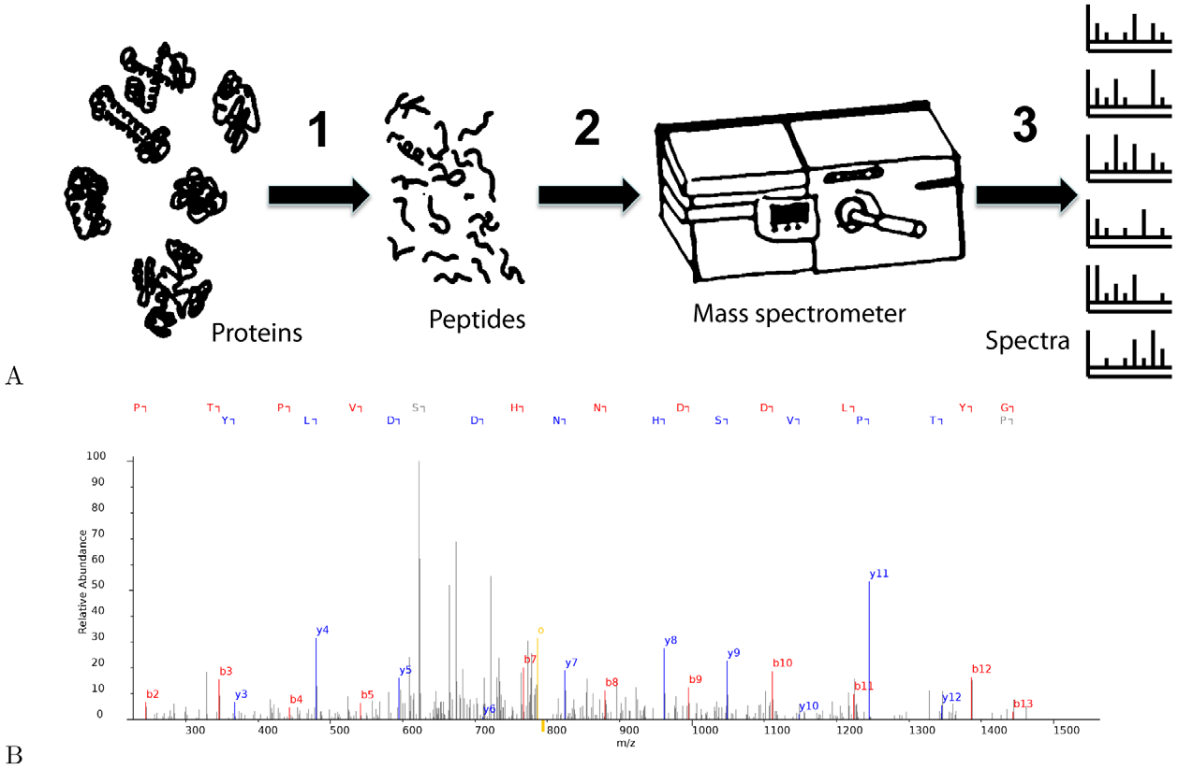


Figure 1. Overview of shotgun proteomics data production. (A) Schematic of a typical shotgun proteomics experiment. The three steps—(1) cleaving proteins into peptides, (2) separation of peptides using liquid chromatography, and (3) tandem mass spectrometry analysis—are described in the text. (B) A sample fragmentation spectrum, along with the peptide responsible for generating the spectrum.
doi:10.1371/journal.pcbi.1002296.g001

third step, which occurs inside the mass spectrometer, involves two rounds of mass spectrometry. Approximately every second, the device analyzes the population of $\sim$20,000 intact peptides that most recently exited from the liquid chromatography column. Then, based on this initial analysis, the machine selects approximately five distinct peptide species for fragmentation. Each of these fragmented species is isolated and subjected to a second round of mass spectrometry analysis. The resulting "fragmentation spectra" are the primary output of the experiment.

A sample fragmentation spectrum is shown in Figure 1B. During the fragmentation process, each amino acid sequence is typically cleaved once, so cleavage of the population results in a variety of observed prefix and suffix sequences. Each of these subpeptides is characterized by its mass-to-charge ratio (m/z, shown on the horizontal axis) and a corresponding intensity (unitless, shown on the vertical axis). The primary analysis challenge is to infer, for each observed fragmentation spectrum, the peptide sequence that was responsible for generating the spectrum.

Peptide and Protein Identification

The Spectrum Identification Problem

The spectrum identification problem is difficult to solve primarily because of noise in the observed spectrum. In general, the x-axis of the observed spectrum is known with relatively high precision. However, in any given spectrum, many expected fragment ions will fail to be observed, and the spectrum is also likely to contain a variety of additional, unexplained peaks. These unexplained peaks may result from unusual fragmentation events, in which small molecular groups are shed from the peptide during fragmentation, or from contaminating molecules (peptides or other small molecules) that are present in the mass spectrometer along with the target peptide species.

In practice, solutions to the spectrum identification problem fall into four general categories. By far the most commonly used approach is database search. The first computer program to use a database search procedure to identify fragmentation spectra was SEQUEST [3], and SEQUEST's basic algorithm (not including the function used to score individual peptide-spectrum matches) is still used by essentially all database search tools available today. The approach is as follows. We are given a spectrum S, a peptide database P, a precursor mass m (i.e., the observed mass of the intact peptide), and a user-specified precursor mass tolerance δ. The algorithm extracts from the database all peptides whose mass lies within the range $[m-\delta, m+\delta]$. These comprise the set of *candidate peptides*

$$C(m,P,\delta)=\{p : p \in P; |m(p)-m| < \delta\}$$

where $m(p)$ is the calculated mass of peptide p. In practice, depending on the size of the peptide database and the precursor mass tolerance, the number of candidate peptides ranges from hundreds to hundreds of thousands. Each candidate peptide p is compared to the observed spectrum using a score function $\Xi(\cdot,\cdot)$. Frequently, the score function generates a theoretical spectrum for the given peptide and then compares the observed and theoretical spectra to one another. The program reports the candidate peptide that scores highest with respect to the observed spectrum:

$$\arg \max_{p \in C(m,P,\delta)} \Xi(S,p).$$

Database search methods differ primarily in their choice of score function.

An alternative to database search is de novo spectrum identification, in which the "database" of candidate peptides consists of the entire universe of possible amino acid sequences. A variety of graph-based dynamic programming methods can efficiently solve this problem, but in practice many spectra do not contain sufficient information to uniquely identify the correct peptide. Consequently, de novo identification methods generally fail to provide as many correct identifications as database search methods. Conversely, of course, de novo approaches are necessary when a peptide database is unavailable—i.e., for analysis of organims whose genomes have not yet been sequenced—or when the user is interested in identifying novel protein isoforms or polymporphisms.

Tag-based methods occupy an appealing middle ground between database search and de novo methods. Here, the basic idea is to use de novo analysis to identify a collection of subpeptides ("tags") that are hypothesized to occur in the sequence, and then extract candidates from a database that contain the tags. Tag-based methods can be quite fast, and retain the ability to partially identify spectra for which the corresponding peptide is not in the database.

Finally, so-called library search methods identify spectra by comparing them to a library of previously identified spectra. These methods suffer a bit from a chicken-and-egg problem, in the sense that you must first somehow identify the spectra that go into the library. However, once you have successfully built such a library, searching an observed spectrum against real spectra is likely to give better results than searching against theoretical spectra. The caveat is that you have to be sure that your library does not contain false positives or chimeric spectra, i.e., spectra that were generated by a heterogeneous population of two or more co-eluting peptides.

Machine learning methods have been applied extensively to the spectrum identification problem, primarily as post-processors that discriminate between correct and incorrect identifications. Using methods such as support vector machines [4], linear discriminant analysis [5], or decision trees [6], these methods can dramatically increase the percentage of spectra from a given experiment that are confidently identified. Particularly powerful are semi-supervised learning methods [7,8] that dynamically adjust their ranking scheme on the basis of characteristics of a given data set.

Protein Identification

Once the peptide responsible for generating each observed spectrum has been identified, the downstream task of deciding which proteins are present in the sample seems like it should involve a straightforward process of aggregating evidence over all the spectra associated with a given protein. Unfortunately, this task is made much more difficult by the presence of so-called degenerate peptides, i.e., peptides that occur in multiple proteins.

Protein identification algorithms have improved significantly over the last decade. Early methods used simple heuristics to identify high-confidence proteins that contain a specified number of high-confidence peptide assignments [9]. A more sophisticated version of this approach employs a parsimony argument and attempts to find a minimal set of proteins that explain the observed identified spectra [10]. The most widely used method is pseudo-probabilistic, employing an expectation-maximization-like procedure to apportion evidence from each degenerate peptide among its corresponding parent proteins [11]. Other, related approaches either handle peptide degeneracy in a similar, heuristic fashion [12–14] or ignore the degeneracy entirely [15,16]. Only recently have several groups proposed algorithms that directly solve the full protein identification problem within a rigorous probabilistic framework [17,18].

Computational and Statistical Challenges

Despite the almost dizzying array of existing methods for identifying peptides and proteins from shotgun proteomics data (reviewed in [19]), many significant analytical challenges remain. Perhaps most obvious is the need for algorithms that successfully identify proteins that are not in the database, either because they are polymorphic or because they contain post-translational modifications. The difficulty here is two-fold: how to make the search efficient, and how to successfully control the rate of false positive identifications, especially in the case when a wide variety of polymorphisms or modifications are allowed.

More fundamentally, the spectrum identification problem could likely benefit from the application of a rich, generative model of the peptide fragmentation process. Several such models have been described in the spectrum identification literature [20,21], but none take into account the relatively rich literature on peptide fragmentation (reviewed in [22]).

A source of ongoing confusion and controversy in the field is the assignment of statistical confidence estimates to spectrum, peptide, or protein identifications. Varying protocols have been proposed, based upon empirical null distributions created by searching the spectra against a "decoy" database of shuffled or reversed peptide sequences [23,24], upon procedures that involve fitting a parametric distribution to the empirical score distribution [25–27], or upon score functions for which exact p-values can be calculated under a simple zero-order Markov null model of peptides [28]. Critical assessment and comparison of many of these methods, especially with respect to potential biases incurred during evaluation of new algorithsm, is lacking in the literature. Furthermore, extending a statistic that is calculated with respect to individual spectra up to the peptide and protein levels is a non-trivial and relatively unexplored subject.

Historically, the development of tools for analyzing shotgun proteomics data has occurred in a stepwise fashion. The result is that small subtasks—assigning charge states to spectra, mapping peptides to spectra, re-ranking and assigning statistical confidence to spectrum identifications, and computing protein-level posteriors—have been solved separately. Clearly, in the long run, this piecemeal approach should be replaced by a joint model in which all relevant aspects of the experiment are taken into account. Such a joint model has the potential to model dependencies among variables at the spectrum, peptide, and protein level that are currently decoupled.

A perhaps equally challenging task is to convey the results of such a rich model to the user in a useful fashion. With current probabilistic protein identification tools, the standard approach of reporting a ranked list of proteins is insufficient. Even when we allow the list to contain groups of redundant proteins (i.e., protein isoforms that are indistinguishable on the basis of the observed spectra), crucial information about dependencies among the protein identifications is lost. For example, it is difficult to convey the information that, e.g., either protein A or protein B was present in the sample, but probably not both.

Authors' Biographies

William Stafford Noble (formerly William Noble Grundy) received the PhD in computer science and cognitive science from University of California (UC) San Diego in 1998. After a one-year postdoc with David Haussler at UC Santa Cruz, he became an Assistant Professor in the Department of Computer Science at Columbia University. In 2002, he joined the faculty of the Department of Genome Sciences at the University of Washington. He is now a Professor of Genome Sciences, with adjunct appointments in the Department of Computer Science and Engineering and the Department of Medicine. His research group develops and applies statistical and machine learning techniques for modeling and understanding biological processes at the molecular level. Dr. Noble is a member of three editorial boards, is the recipient of National Science Foundation CAREER award, and is a Sloan Research Fellow.

Michael J. MacCoss has been working with mass spectrometry instrumentation since 1994 when he was an undergraduate in a stable isotope geochemistry lab at the University of Vermont. He became interested in biomedical applications working in Dr. Patrick Griffin's protein mass spectrometry lab at Merck Research Laboratories during two summer internships in 1995 and 1996. In 2001, he completed a PhD in analytical chemistry with Professor Dwight Matthews in the development of stable isotope and mass spectrometry methodologies for the measurement of human amino acid and protein metabolism. After completing his degree, Dr. MacCoss moved to The Scripps Research Institute to work with Professor John R. Yates III as a postdoctoral fellow. Dr. MacCoss moved to the University of Washington in 2004 as an Assistant Professor of Genome Sciences, where his lab has focused on the development and application of mass spectrometry–based technologies for the high-throughput characterization of complex protein mixtures. In 2009 he was promoted to Associate Professor.

Another area where proteomics is in great need of guidance is experimental design. Proteomics experiments can be expensive and time consuming, yet most are being pursued without proper care in minimizing batch or systematic effects. Furthermore, most proteomics experiments are not powered correctly. Like gene expression analyses previouly, proteomics practices could benefit greatly from lessons learned in experimental design from classical statistics.

Protein Quantification

The next logical step, after developing methods to *identify* the proteins in a complex mixture, is to develop methods to *quantify* the proteins. Quantification provides a more complete picture of the molecular contents of the cell, and allows us to generate or test specific hypotheses regarding the relationship between protein abundance and fundamental cellular processes or disease states.

Existing methods for protein quantification fall into three categories. *Stable isotope labeling* methods [29] perform relative protein quantification by incorporating a distinct heavy isotope tag to a sample to use as an internal standard. This labeled internal standard is then mixed into one or more other samples, and the relative signal intensity of the peptide measured by the mass spectrometer is compared to the measured intensity of the same peptide containing the heavy stable isotope label. This type of approach is quite powerful but the isotopic labeling step imposes significant overhead and limits the general applicability of these methods. *Spectral counting* methods [30–33] rely on counting the number of spectra that map to a given protein across multiple experiments. Spectral counting methods are not very accurate, because these methods fail to take into account the data-dependent acquisition that leads to the selection of peptides for fragmentation. However, spectral counting is appealing because the required counts are relatively easy to compute. Finally, *peptide chromatographic peak intensity* methods [34–37] use the area under the precursor ion peak as a proxy for peptide abundance. In contrast to spectral counting, methods based on peak areas are potentially much more accurate, but these methods require highly reproducible liquid chromatography as well as accurate methods for chromatographic alignment and identification of relevant spectral features.

Most of the mass spectrometry quantification literature focuses on measuring the relative abundance of the same protein across different samples. This allows, for example, comparing protein expression across different patients, tissues, or developmental stages. More difficult is estimating the relative abundance of two proteins within the same sample. For this task, the most commonly used, low-throughput approach is to calibrate the response of an individual peptide in a targeted selective reaction monitoring method (described in the next section). For large-scale quantification of many proteins, a few methods use a standard shotgun proteomics experiment, and either use a simple learning procedure [30] or rely on the observation that the three most intense peptides ionize similarly between proteins [38,39].

Targeted Proteomics

Many high-throughput proteomics experiments aim to identify or quantify all proteins in a complex sample. In contrast, selected reaction monitoring (SRM) experiments [40,41] seek to quantify a smaller, specified set of proteins, e.g., a panel of biomarkers or members of a pathway of interest.

In an SRM experiment, the mass spectrometer is set to monitor the m/z values for a small number of peptides, as well as a specific fragment ion for each peptide. Each m/z pair, corresponding to the intact peptide and its fragment ion, is called a *transition*. Monitoring a small number of transitions, rather than scanning the entire m/z range, yields dramatically increased sensitivity relative to conventional "full scan" techniques. The goal of SRM is to select transitions that best detect the proteins of interest, subject to the constraint that such experiments can monitor only $\sim$1,000 transitions per run in an automated fashion [42]. Thus, two complementary criteria must be optimized: (1) detect as many proteins of interest as possible and (2) accurately estimate the abundance of the monitored proteins.

Existing SRM pipelines typically focus on so-called proteotypic peptides [43] that can be easily observed in a mass spectometry experiment and that uniquely identify a specific protein. A variety of methods exist for identifying proteotypic peptides, either based on empirical rules [44,45] or machine learning methods [43,46,47]. After identifying these peptides, SRM protocols typically use them as independent protein identifiers [48], with each peptide contributing equally to the evidence for that protein. What is missing is a method to search for a panel of peptides that jointly provide high quality quantification information about all of the target proteins.

Outlook

Traditionally, most proteomics analysis has been carried out using relatively inexpensive ion trap instruments, which offer fairly

low precision and accuracy on the m/z axis. Higher resolution instruments, which achieve precision of <10 ppm, were expensive and hence more rare. However, with the introduction of *Orbitrap* and improved quadrupole time-of-flight mass spectrometers [49], high-resolution instruments have become much more commonplace. Corresponding analytical methods that fully exploit the information available from high-resolution spectra have not used these data to their full potential.

In a mass spectrometer, intact peptides are characterized in an initial scan, followed by a series of secondary scans that characterize fragmented versions of the same peptides. The most common means of fragmenting peptides between the two scans is collision-induced dissociation, in which the charged peptides collide with neutral molecules such as helium, nitrogen, or argon. However, a variety of other fragmentation methods have been developed, including electron-capture dissociation, electron-transfer dissociation, and infrared multiphoton dissociation [50], and these can provide fragmentation spectra with quite different properties. More recently, some protocols have been developed to alternate between different types of fragmentation methods, with the aim of observing two complementary spectra representing each peptide [51]. Most existing analysis pipelines are tuned to handle one or two of the resulting types of spectra. Although some search engines allow users to select which types of fragmentation ions are included in the search, and some progress has been made recently toward developing score functions that can be adapted to various types of spectra [52], the field is still missing a generic analysis platform that can be adapted automatically and in a principled fashion to handle spectra produced by any given fragmentation protocol.

An interesting consideration is that as the scan speeds of tandem mass spectrometers increase, the difference between targeted and discovery proteomics will become more and more blurred. As an instrument becomes capable of collecting MS/MS spectra continuously across the chromatographic time-scale on an increasingly larger number of peptide precursors, the "m/z space" that remains unsampled will become less significant. Approaches that improve sampling by multiplexing the collection of fragmentation data are promising, but require methods to deconvolve the resulting mixed spectra. Such approaches fall under a general category known as *data-independent acquisition* because they collect their fragmentation data independently of whether a signal is observed within a particular precursor m/z window. Some initial attempts to interpret these data have shoe-horned traditional proteomics analysis pipelines to handle these data [53,54]. Other workflows that have been developed specifically for these unique data have remained proprietary and, thus, not attracted much effort from the academic community to build and improve on these analyses.

One challenge that remains is enabling computational and experimental scientists pursuing proteomics to interact effectively. Frequently, experimentalists design new experiments using suboptimal analysis tools because they do not have the skills or knowledge to pursue alternatives. Likewise, computational scientists can expend a large amount of energy developing solutions to problems that are not interesting to experimentalists. The challenge remains how to get scientists from different disciplines that are based on different cultures and who speak different scientific languages to communicate and collaborate effectively. The RECOMB Satellite Conference on Computational Proteomics is one attempt to solve this problem. Speakers are brought from both the computational and experimental backgrounds. This meeting provides a forum for experimentalists to present problems that they are facing and for computational scientists to present algorithmic and statistical approaches that they have been developing. Ultimately, we need more meetings that foster collaboration between disciplines.

References

1. Liepe J, Mishto M, Textoris-Taube K, Janke K, Keller C, et al. (2010) The 20s proteasome splicing activity discovered by splicemet. PLoS Comput Biol 6: e1000830. doi:10.1371/journal.pcbi.1000830.
2. Marchese R, Grandori R, Carloni P, Raugei S (2010) On the zwitterionic nature of gas-phase peptides and protein ions. PLoS Comput Biol 6: e1000775. doi:10.1371/journal.pcbi.1000775.
3. Eng JK, McCormack AL, Yates III JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Mass Spectrom 5: 976–989.
4. Anderson DC, Li W, Payan DG, Noble WS (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and sequest scores. J Proteome Res 2: 137–146.
5. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. Anal Chem 74: 5383–5392.
6. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22: 214–219.
7. Käll L, Canterbury J, Weston J, Noble WS, MacCoss MJ (2007) A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. Nat Methods 4: 923–25.
8. Choi H, Nesvizhskii AI (2008) Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 7: 254–265.
9. Tabb DL, McDonald WH, Yates III JR (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res 1: 21–26.
10. Ma ZQ, Dasari S, Chambers MC, Litton M, Sobecki SM, et al. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 8: 3872–3881.
11. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646–4658.
12. Searle BC (2010) Scaffold: A bioinformatic tool for validating ms/ms-based proteomic studies. Proteomics 10: 1265–1269.
13. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, et al. (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteomics 6: 527–536.
14. Feng J, Naiman DQ, Cooper B (2007) Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. Bioinformatics 23: 2210–2217.
15. Shen C, Wang Z, Shankar G, Zhang X, Li L (2008) A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. Bioinformatics 24: 202–208.
16. Li Q, MacCoss MJ, Stephens M (2010) A nested mixture model for protein identification using mass spectrometry. Annals of Applied Sciences 4: 962–987.
17. Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, et al. (2008) A Bayesian approach to protein inference problem in shotgun proteomics. In: Vingron M, Wong L, eds. Proceedings of the Twelfth Annual International Conference on Computational Molecular Biology. Berlin: Springer, Volume 12 of Lecture Notes in Bioinformatics. pp 167–180.
18. Serang O, MacCoss MJ, Noble WS (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res 9: 5346–5357.
19. Nezvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of Proteomics 73: 2092–2123.
20. Wan Y, Chen T (2005) PepHMM: A hidden Markov model based scoring function for mass spectrometry database search. Anal Chem 78: 432–437.
21. Klammer AA, Reynolds SR, Hoopmann M, MacCoss MJ, Bilmes J, et al. (2008) Modeling peptide fragmentation with dynamic Bayesian networks yields improved tandem mass spectrum identification. Bioinformatics 24: i348–i356.
22. Barton SJ, Whittaker JC (2009) Review of factors that inflence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. Mass Spectrom Rev 28: 177–187.
23. Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7: 29–34.
24. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4: 207–214.

25. Fenyo D, Beavis RC (2003) A method for assessing the statistical significance of mass spectrometrybased protein identification using general scoring schemes. Anal Chem 75: 768–774.
26. Klammer AA, Park CY, Noble WS (2009) Statistical calibration of the sequest XCorr function. J Proteome Res 8: 2106–2113.
27. Spirin V, Shpunt A, Seebacher J, Gentzel M, Shevchenko A, et al. (2011) Assigning spectrum-specific p-values to protein identifications by mass spectrometry. Bioinformatics 27: 1128–1134.
28. Kim S, Gupta N, Pevzner PA (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J Proteome Res 7: 3354–3363.
29. Gevaert K, Impens F, Ghesquière B, Van Damme P, Lambrechts A, et al. (2008) Stable isotopic labeling in proteomics. Proteomics 8: 4873–4885.
30. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2006) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117–124.
31. Griffin NM, Yu J, Long F, Oh P, Shore S, et al. (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol 28: 83–89.
32. Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, et al. (2006) Quantitative proteomic analysis of distinct mammalian mediator complexes using normalized spectral abundance factors. Proc Natl Acad Sci U S A 103: 18928–18933.
33. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, et al. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics 4: 1487–1502.
34. Clough T, Key M, Ott I, Ragg S, Schadow G, et al. (2009) Protein quantification in label-free LC-MS experiments. J Proteome Res 8: 5275–5284.
35. Karpievitch Y, Stanley J, Taverner T, Huang J, Adkins JN, et al. (2009) A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics 25: 2028–2034.
36. Bondarenko PV, Chelius D, Shaler TA (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. Anal Chem 74: 4741–4749.
37. Roy SM, Becker CH (2007) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling. Methods Mol Biol 359: 87–105.
38. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics 5: 144–156.
39. Malmstrom J, Beck M, Schmidt A, Lange V, Deutsch EW, et al. (2009) Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. Nature 460: 762–765.
40. Kuhn E, Wu J, Karl J, Liao H, Zolg W, et al. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. Proteomics 4: 1175–1186.
41. Lange V, Picotti P, Domon B, Aebersold RH (2008) Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 4: 222.
42. Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, et al. (2008) Targeted quantitative analysis of *streptococcus pyogenes* virulence factors by multiple reaction monitoring. Mol Cell Proteomics 7: 1489–1500.
43. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, et al. (2006) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25: 125–131.
44. Bronstrup M (2004) Absolute quantification strategies in proteomics based on mass spectrometry. Expert Reviews of Proteomics 1: 503–512.
45. Kirkpatrick DS, Gerber SA, Gygi SP (2005) The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. Methods 35: 265–273.
46. Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, et al. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics 22: e481–e488.
47. Webb-Robertson BJM, Cannon WR, Oehmen CS, Shah AR, G V, et al. (2008) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. Bioinformatics 24: 1503–1509.
48. Kuster B, Schirle M, Mallick P, Aebersold RH (2005) Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 6: 577–583.
49. Olsen JV, de Godoy LMF, Li G, Macek B, Mortensen P, et al. (2005) Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics 4: 2010–2021.
50. Sleno L, Volmer DA (2004) Ion activation methods for tandem mass spectrometry. J Mass Spectrom 39: 1091–1112.
51. Swaney DL, McAlister GC, Coon JC (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat Methods 5: 959–964.
52. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, et al. (2010) The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol Cell Proteomics 9: 2840–2852.
53. Venable JD, Dong MQ, Wohlsclegel J, Dillin A, Yates III JR (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 1: 39–45.
54. Panchaud A, Jung S, Shaffer SA, Aitchison JD, Goodlett DR (2011) Faster, quantitative, and accurate precursor acquisition independent from ion count. Anal Chem 83: 2250–2257.

D.9 Protein structure

Further explanations for Chapter 9.

D.9.1 CASP: a blind test of structure prediction

- Elofsson (2023) — A short review of what CASP15 showed after AlphaFold2: where monomer prediction stands, the progress for multimers and the categories that remain difficult. Useful as a compact account of how CASP results are read. *(review; CC BY 4.0)*

  Progress at protein structure prediction, as seen in CASP15. Elofsson (2023). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2023.102594

  > In Dec 2020, the results of AlphaFold version 2 were presented at CASP14, sparking a revolution in the field of protein structure predictions. For the first time, a purely computational method could challenge experimental accuracy for structure prediction of single protein domains. The code of AlphaFold v2 was released in the summer of 2021, and since then, it has been shown that it can be used to accurately predict the structure of most ordered proteins and many protein-protein interactions. It has also sparked an explosion of development in the field, improving AI-based methods to predict protein complexes, disordered regions, and protein design. Here I will review some of the inventions sparked by the release of AlphaFold.

- Abriata and Dal Peraro (2025) — Summarises the CASP16 results from the point of view of a user: which kinds of targets (monomers, complexes, ligands, nucleic acids) can be modelled reliably and which cannot. Helps to translate CASP metrics into practical expectations. *(review; CC BY 4.0)*

  Practical Outcomes From CASP16 for Users in Need of Biomolecular Structure Prediction. Abriata & Dal Peraro (2025). *Proteins: Structure, Function, and Bioinformatics*. doi:10.1002/prot.70078

  > The 16th Critical Assessment of Structure Prediction benchmarked advancements in biomolecular modeling, particularly in the context of AlphaFold 2 and 3 systems. Protein monomer and domain prediction is largely solved, with barely any space for further improvements at the backbone level except for very specific details, irregular secondary structures, and mutational effects that remain challenging to predict. For protein assemblies, AF-based methods, especially when expertly guided or enhanced by servers like those from the Yang, Zheng/Zhang, and Cheng lab, show progress, though complex topologies and in particular antibody-antigen interactions are still difficult. Notably, a priori knowledge of stoichiometry significantly aids assembly prediction. Protein-ligand co-folding with AF3 demonstrated strong potential for pose prediction, outperforming many participants and some dedicated docking tools in baseline tests, but several caveats hold as discussed. Ligand affinity prediction is totally unreliable. Nucleic acid structure prediction lags considerably, heavily relying on 3D templates and expert human intervention, even AF3 showing substantial limitations. Overall, on all fronts, AF3's modeling capabilities are at or close to the state of the art; additionally, it shows slight improvements over AF2 and more detailed confidence metrics than it. We guide users on tool selection, realistic accuracy expectations, and persistent challenges, emphasizing the critical role of confidence metrics in interpreting AI-generated models.

D.9.2 Background: amino acids, energetics and enzymes

- Jakubowski & Flatt (2024) — Open textbook chapter on the chemistry of the amino acids, side-chain properties and the peptide bond; the neighbouring chapters of the same book cover protein structure and enzyme activity. Good for revising the biochemistry this section assumes. *(book chapter; CC BY-NC-SA 4.0)*

  Amino Acids, Peptides, and Proteins (Chapter 3 of Fundamentals of Biochemistry). Jakubowski & Flatt (2024). *LibreTexts Biology (open textbook)*. https://bio.libretexts.org/Bookshelves/Biochemistry/Fundamentals_of_Biochemistry_(Jakubowski_and_Flatt)/01%3A_Unit_I-_Structure_and_Catalysis/03%3A_Amino_Acids_Peptides_and_Proteins

- Jarmoskaite et al. (2020) — Explains what a dissociation constant means and the controls needed to measure it correctly, with many examples of common errors. Gives a concrete feel for the thermodynamics of binding discussed in the section. *(tutorial; CC BY 4.0)*

  How to measure and evaluate binding affinities. Jarmoskaite et al. (2020). *eLife*. doi:10.7554/elife.57264

  > Quantitative measurements of biomolecule associations are central to biological understanding and are needed to build and test predictive and mechanistic models. Given the advances in high-throughput technologies and the projected increase in the availability of binding data, we found it especially timely to evaluate the current standards for performing and reporting binding measurements. A review of 100 studies revealed that in most cases essential controls for establishing the appropriate incubation time and concentration regime were not documented, making it impossible to determine measurement reliability. Moreover, several reported affinities could be concluded to be incorrect, thereby impacting biological interpretations. Given these challenges, we provide a framework for a broad range of researchers to evaluate, teach about, perform, and clearly document high-quality equilibrium binding measurements. We apply this framework and explain underlying fundamental concepts through experimental examples with the RNA-binding protein Puf4.

- Cornish-Bowden (2015) — A short historical and conceptual account of the Michaelis–Menten equation, what kcat and Km mean and how they are often misinterpreted. Useful background for the enzyme part of the section. *(perspective; CC BY 4.0)*

  One hundred years of Michaelis–Menten kinetics. Cornish-Bowden (2015). *Perspectives in Science*. doi:10.1016/j.pisc.2014.12.002

  > The year 2013 marked the centenary of the paper of Leonor Michaelis and Maud Menten (Michaelis and Menten, 1913), and the 110th anniversary of the doctoral thesis of Victor Henri (Henri, 1903). These publications have had an enormous influence on the progress of biochemistry, and are more often cited in the 21st century than they were in the 20th. Henri laid the groundwork for the understanding of enzyme mechanisms, but his experimental design was open to criticism. He reached essentially correct conclusions about the action of invertase, but he took no steps to control the hydrogen-ion concentration, and he took no account of the spontaneous mutarotation of the glucose produced in the reaction. Michaelis and Menten corrected these shortcomings, and in addition they introduced the initial-rate method of analysis, which has proved much simpler to apply than the methods based on time courses that it replaced. In this way they defined the methodology for steady-state experiments that has remained standard for 100 years.

D.9.3 Protein structure and how it is determined

- Vénien-Bryan et al. (2017) — Compares the principles, strengths and limitations of X-ray crystallography and single-particle cryo-EM, written for readers who are not specialists in either. Helps to understand why resolution and sample requirements differ between the methods. *(review; CC BY 2.0 UK)*

  Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. Vénien-Bryan et al. (2017). *Acta Crystallographica Section F Structural Biology Communications*. doi:10.1107/s2053230x17003740

  > The invention of the electron microscope has greatly enhanced the view scientists have of small structural details. Since its implementation, this technology has undergone considerable evolution and the resolution that can be obtained for biological objects has been extended. In addition, the latest generation of cryo-electron microscopes equipped with direct electron detectors and software for the automated collection of images, in combination with the use of advanced image-analysis methods, has dramatically improved the performance of this technique in terms of resolution. While calculating a sub-10 Å resolution structure was an accomplishment less than a decade ago, it is now common to generate structures at sub-5 Å resolution and even better. It is becoming possible to relatively quickly obtain high-resolution structures of biological molecules, in particular large ones (>500 kDa) which, in some cases, have resisted more conventional methods such as X-ray crystallography or nuclear magnetic resonance (NMR). Such newly resolved structures may, for the first time, shed light on the precise mechanisms that are essential for cellular physiological processes. The ability to attain atomic resolution may support the development

> of new drugs that target these proteins, allowing medicinal chemists to understand the intimacy of the relationship between their molecules and targets. In addition, recent developments in cryo-electron microscopy combined with image analysis can provide unique information on the conformational variability of macromolecular complexes. Conformational flexibility of macromolecular complexes can be investigated using cryo-electron microscopy and multiconformation reconstruction methods. However, the biochemical quality of the sample remains the major bottleneck to routine cryo-electron microscopy-based determination of structures at very high resolution.

- Laurents (2022) — Reviews what NMR spectroscopy measures (structure, dynamics, disorder) and how it complements AlphaFold2 models. Useful for the NMR subsection and for seeing what a static prediction does not capture. *(review; CC BY 4.0)*

  AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function. Laurents (2022). *Frontiers in Molecular Biosciences*. doi:10.3389/fmolb.2022.906437

  > The artificial intelligence program AlphaFold 2 is revolutionizing the field of protein structure determination as it accurately predicts the 3D structure of two thirds of the human proteome. Its predictions can be used directly as structural models or indirectly as aids for experimental structure determination using X-ray crystallography, CryoEM or NMR spectroscopy. Nevertheless, AlphaFold 2 can neither afford insight into how proteins fold, nor can it determine protein stability or dynamics. Rare folds or minor alternative conformations are also not predicted by AlphaFold 2 and the program does not forecast the impact of post translational modifications, mutations or ligand binding. The remaining third of human proteome which is poorly predicted largely corresponds to intrinsically disordered regions of proteins. Key to regulation and signaling networks, these disordered regions often form biomolecular condensates or amyloids. Fortunately, the limitations of AlphaFold 2 are largely complemented by NMR spectroscopy. This experimental approach provides information on protein folding and dynamics as well as biomolecular condensates and amyloids and their modulation by experimental conditions, small molecules, post translational modifications, mutations, flanking sequence, interactions with other proteins, RNA and virus. Together, NMR spectroscopy and AlphaFold 2 can collaborate to advance our comprehension of proteins.

- Terwilliger et al. (2023) — Compares AlphaFold models directly with experimental electron-density maps and shows where even high-confidence predictions differ from the data. Directly relevant to the subsection on experimental structures versus predicted models. *(perspective; CC BY 4.0)*

  AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Terwilliger et al. (2023). *Nature Methods*. doi:10.1038/s41592-023-02087-4

  > Artificial intelligence-based protein structure prediction methods such as AlphaFold have revolutionized structural biology. The accuracies of these predictions vary, however, and they do not take into account ligands, covalent modifications or other environmental factors. Here, we evaluate how well AlphaFold predictions can be expected to describe the structure of a protein by comparing predictions directly with experimental crystallographic maps. In many cases, AlphaFold predictions matched experimental maps remarkably closely. In other cases, even very high-confidence predictions differed from experimental maps on a global scale through distortion and domain orientation, and on a local scale in backbone and side-chain conformation. We suggest considering AlphaFold predictions as exceptionally useful hypotheses. We further suggest that it is important to consider the confidence in prediction when interpreting AlphaFold predictions and to carry out experimental structure determination to verify structural details, particularly those that involve interactions not included in the prediction.

D.9.4 Protein folding

- Gianni et al. (2021) — Introduces the energy-landscape view of folding, the principle of minimal frustration and how residual frustration relates to function and disorder. A readable route from folding funnels to current ideas. *(review; CC BY 4.0)*

  Fuzziness and Frustration in the Energy Landscape of Protein Folding, Function, and Assembly. Gianni et al. (2021). *Accounts of Chemical Research*. doi:10.1021/acs.accounts.0c00813

> Are all protein interactions fully optimized? Do suboptimal interactions compromise specificity? What is the functional impact of frustration? Why does evolution not optimize some contacts? Proteins and their complexes are best described as ensembles of states populating an energy landscape. These ensembles vary in breadth from narrow ensembles clustered around a single average X-ray structure to broader ensembles encompassing a few different functional "taxonomic" states on to near continua of rapidly interconverting conformations, which are called "fuzzy" or even "intrinsically disordered". Here we aim to provide a comprehensive framework for confronting the structural and dynamical continuum of protein assemblies by combining the concepts of energetic frustration and interaction fuzziness. The diversity of the protein structural ensemble arises from the frustrated conflicts between the interactions that create the energy landscape. When frustration is minimal after folding, it results in a narrow ensemble, but residual frustrated interactions result in fuzzy ensembles, and this fuzziness allows a versatile repertoire of biological interactions. Here we discuss how fuzziness and frustration play off each other as proteins fold and assemble, viewing their significance from energetic, functional, and evolutionary perspectives.We demonstrate, in particular, that the common physical origin of both concepts is related to the ruggedness of the energy landscapes, intramolecular in the case of frustration and intermolecular in the case of fuzziness. Within this framework, we show that alternative sets of suboptimal contacts may encode specificity without achieving a single structural optimum. Thus, we demonstrate that structured complexes may not be optimized, and energetic frustration is realized via different sets of contacts leading to multiplicity of specific complexes. Furthermore, we propose that these suboptimal, frustrated, or fuzzy interactions are under evolutionary selection and expand the biological repertoire by providing a multiplicity of biological activities. In accord, we show that non-native interactions in folding or interaction landscapes can cooperate to generate diverse functional states, which are essential to facilitate adaptation to different cellular conditions. Thus, we propose that not fully optimized structures may actually be beneficial for biological activities of proteins via an alternative set of suboptimal interactions. The importance of such variability has not been recognized across different areas of biology.This account provides a modern view on folding, function, and assembly across the protein universe. The physical framework presented here is applicable to the structure and dynamics continuum of proteins and opens up new perspectives for drug design involving not fully structured, highly dynamic protein assemblies.

- Bowman (2024) — A short review arguing that predicting a single structure does not solve the folding problem, and that folding pathways and conformational ensembles remain open questions. Fits the subsection on folding and prediction. *(review; CC BY 4.0)*

  AlphaFold and Protein Folding: Not Dead Yet! The Frontier Is Conformational Ensembles. Bowman (2024). *Annual Review of Biomedical Data Science*. doi:10.1146/annurev-biodatasci-102423-011435

  > Like the black knight in the classic Monty Python movie, grand scientific challenges such as protein folding are hard to finish off. Notably, AlphaFold is revolutionizing structural biology by bringing highly accurate structure prediction to the masses and opening up innumerable new avenues of research. Despite this enormous success, calling structure prediction, much less protein folding and related problems, "solved" is dangerous, as doing so could stymie further progress. Imagine what the world would be like if we had declared flight solved after the first commercial airlines opened and stopped investing in further research and development. Likewise, there are still important limitations to structure prediction that we would benefit from addressing. Moreover, we are limited in our understanding of the enormous diversity of different structures a single protein can adopt (called a conformational ensemble) and the dynamics by which a protein explores this space. What is clear is that conformational ensembles are critical to protein function, and understanding this aspect of protein dynamics will advance our ability to design new proteins and drugs.

D.9.5 Structure comparison and evolution

- Sahakyan et al. (2025) — A brief review of methods for detecting distant relationships between proteins, from sequence profiles to structure comparison with large predicted-structure databases. Shows why structure comparison reaches further back in evolution than sequence comparison. *(review; CC BY 4.0)*

  Exploring the protein universe with distant similarity detection methods. Sahakyan et al. (2025). *Protein Science*. doi:10.1002/pro.70397

> During the last few years, the body of data on proteins is expanding almost exponentially with the development of advanced methods for gene sequencing, protein structure determination, particularly by cryoelectron microscopy, and structure prediction using artificial intelligence-based approaches. These developments create the potential for a comprehensive exploration of the protein universe, the entirety of the proteins existing in the biosphere. Elucidation of the relationships among proteins including the most distant ones, where only the core fold is shared, is crucial for understanding protein functions, folding mechanisms, and evolution, as well as the evolution of cellular life forms and viruses. In this brief review, we discuss methods that shaped the field of protein bioinformatics, first, through comparative sequence analysis, and the recent developments in protein structure prediction that transformed the state of the art in the comparative analysis of distantly related proteins. The combination of the rapidly growing databases of genome and metagenome sequences with sensitive methods for sequence comparison and the new generation of structure analysis tools can make charting the protein universe at the structural level a realistic goal.

- Kulkarni et al. (2025) — A perspective on how views of the protein structure universe have changed, including fold switching, disorder and amyloid. Useful as background for what structure comparison can and cannot reveal about evolution. *(perspective; CC BY 4.0)*

  Evolving concepts of the protein universe. Kulkarni et al. (2025). *iScience*. doi:10.1016/j.isci.2025.112012

  > The protein universe is the collection of all proteins on earth from all organisms both extant and extinct. Classical studies on protein folding suggested that proteins exist as a unique three-dimensional conformation that is dictated by the genetic code and is critical for function. In this perspective, we discuss ideas and developments that emerged over the past three decades regarding the protein structure-function paradigm. It is now clear that ordered (active/functional) and disordered/denatured (and hence inactive/non-functional) represent a continuum of states rather than binary states. Some proteins can switch folds without sequence change. Others exist as conformational ensembles lacking defined structure yet play critical roles in many biological processes, including forming membrane-less organelles driven by liquid-liquid phase separation. Numerous diverse proteins harbor segments with the potential to form amyloid fibrils, many of which are functional, and some possess prion-like properties enabling conformation-based transfer of heritable information. Taken together, these developments reveal the remarkable complexity of the protein universe.

D.9.6 Protein structure databases

- Burley et al. (2022) — Describes what the PDB contains, how entries are deposited and validated, and how the archive is used. A good overview of the central structure database. *(review; CC BY 4.0)*

  Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. Burley et al. (2022). *Biomolecules*. doi:10.3390/biom12101425

  > The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), funded by the United States National Science Foundation, National Institutes of Health, and Department of Energy, supports structural biologists and Protein Data Bank (PDB) data users around the world. The RCSB PDB, a founding member of the Worldwide Protein Data Bank (wwPDB) partnership, serves as the US data center for the global PDB archive housing experimentally-determined three-dimensional (3D) structure data for biological macromolecules. As the wwPDB-designated Archive Keeper, RCSB PDB is also responsible for the security of PDB data and weekly update of the archive. RCSB PDB serves tens of thousands of data depositors (using macromolecular crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, and micro-electron diffraction) annually working on all permanently inhabited continents. RCSB PDB makes PDB data available from its research-focused web portal at no charge and without usage restrictions to many millions of PDB data consumers around the globe. It also provides educators, students, and the general public with an introduction to the PDB and related training materials through its outreach and education-focused web portal. This review article describes growth of the PDB, examines evolution of experimental methods for structure determination viewed through the lens of the PDB archive, and provides a detailed accounting of PDB archival holdings and their utilization by researchers, educators, and students worldwide.

- Fleming et al. (2025) — Describes the contents and use of the AlphaFold Protein Structure Database and the 3D-Beacons network that links predicted and experimental models. Useful for understanding how predicted structures are distributed and accessed. *(review; CC BY 4.0)*

  AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. Fleming et al. (2025). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2025.168967

  > The AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) has made significant strides in enhancing its utility and accessibility for the life science research community. The recent integration of AlphaMissense predictions enables access to the pathogenicity of human protein missense variants, with an innovative and interactive heatmap and 3D visualisation that display variant data at the residue level. Users can now toggle between structure model quality (pLDDT) and average pathogenicity scores, providing insights into the implications of specific residue changes. The Foldseek integration offers a rapid and accurate method for protein structure searches and comparisons. Bulk data download options further facilitate comprehensive data analysis and integration with other computational tools. The 3D-Beacons framework (https://www.ebi.ac.uk/pdbe/pdbe-kb/3dbeacons/) has also been enhanced with detailed annotation endpoints (such as AlphaMissense data) and integrates LevyLab's dataset of homomeric AlphaFold 2 models. These advancements significantly improve the functionality and accessibility of these resources, enabling discoveries using structure data.

- Waman et al. (2024) — Explains the CATH hierarchy (class, architecture, topology, homologous superfamily) and how it has been extended with predicted models. Illustrates how structure classification works in practice. *(review; CC BY 4.0)*

  CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. Waman et al. (2024). *Journal of Molecular Biology*. doi:10.1016/j.jmb.2024.168551

  > CATH (https://www.cathdb.info) classifies domain structures from experimental protein structures in the PDB and predicted structures in the AlphaFold Database (AFDB). To cope with the scale of the predicted data a new NextFlow workflow (CATH-AlphaFlow), has been developed to classify high-quality domains into CATH superfamilies and identify novel fold groups and superfamilies. CATH-AlphaFlow uses a novel state-of-the-art structure-based domain boundary prediction method (ChainSaw) for identifying domains in multi-domain proteins. We applied CATH-AlphaFlow to process PDB structures not classified in CATH and AFDB structures from 21 model organisms, expanding CATH by over 100%. Domains not classified in existing CATH superfamilies or fold groups were used to seed novel folds, giving 253 new folds from PDB structures (September 2023 release) and 96 from AFDB structures of proteomes of 21 model organisms. Where possible, functional annotations were obtained using (i) predictions from publicly available methods (ii) annotations from structural relatives in AFDB/UniProt50. We also predicted functional sites and highly conserved residues. Some folds are associated with important functions such as photosynthetic acclimation (in flowering plants), iron permease activity (in fungi) and post-natal spermatogenesis (in mice). CATH-AlphaFlow will allow us to identify many more CATH relatives in the AFDB, further characterising the protein structure landscape.

D.9.7 Protein structure prediction

- Bertoline et al. (2023) — Covers the history of structure prediction from homology modelling, threading and fragment assembly through coevolution to AlphaFold2. Follows the same order as the section and is a good overview for revision. *(review; CC BY 4.0)*

  Before and after AlphaFold2: An overview of protein structure prediction. Bertoline et al. (2023). *Frontiers in Bioinformatics*. doi:10.3389/fbinf.2023.1120370

  > Three-dimensional protein structure is directly correlated with its function and its determination is critical to understanding biological processes and addressing human health and life science problems in general. Although new protein structures are experimentally obtained over time, there is still a large difference between the number of protein sequences placed in Uniprot and those with resolved tertiary structure. In this context, studies have emerged to predict protein structures by methods based on a template or free modeling. In the last years, different methods have been combined to overcome their individual limitations, until the emergence of AlphaFold2, which demonstrated that predicting protein structure with

> high accuracy at unprecedented scale is possible. Despite its current impact in the field, AlphaFold2 has limitations. Recently, new methods based on protein language models have promised to revolutionize the protein structural biology allowing the discovery of protein structure and function only from evolutionary patterns present on protein sequence. Even though these methods do not reach AlphaFold2 accuracy, they already covered some of its limitations, being able to predict with high accuracy more than 200 million proteins from metagenomic databases. In this mini-review, we provide an overview of the breakthroughs in protein structure prediction before and after AlphaFold2 emergence.

- Kovalevskiy et al. (2024) — Reviews how AlphaFold2 predictions have been validated against experiments and how they have been used. Written by DeepMind authors, so it should be read alongside independent assessments; not peer reviewed. *(review; CC BY 4.0)*

  AlphaFold two years on: validation and impact. Kovalevskiy et al. (2024). doi:10.48550/arXiv.2403.02124

  > Two years on from the initial release of AlphaFold2 we have seen its widespread adoption as a structure prediction tool. Here we discuss some of the latest work based on AlphaFold2, with a particular focus on its use within the structural biology community. This encompasses use cases like speeding up structure determination itself, enabling new computational studies, and building new tools and workflows. We also look at the ongoing validation of AlphaFold2, as its predictions continue to be compared against large numbers of experimental structures to further delineate the model's capabilities and limitations.

- Rennie and Oliver (2025) — Discusses how to interpret predicted models and extends to complexes, ligands and other biomolecules, and proposes guidelines for reporting AlphaFold predictions. Relevant to the subsections on co-folding, limitations and checking a model. *(review; CC BY 4.0)*

  Emerging frontiers in protein structure prediction following the AlphaFold revolution. Rennie & Oliver (2025). *Journal of The Royal Society Interface*. doi:10.1098/rsif.2024.0886

  > Models of protein structures enable molecular understanding of biological processes. Current protein structure prediction tools lie at the interface of biology, chemistry and computer science. Millions of protein structure models have been generated in a very short space of time through a revolution in protein structure prediction driven by deep learning, led by AlphaFold. This has provided a wealth of new structural information. Interpreting these predictions is critical to determining where and when this information is useful. But proteins are not static nor do they act alone, and structures of proteins interacting with other proteins and other biomolecules are critical to a complete understanding of their biological function at the molecular level. This review focuses on the application of state-of-the-art protein structure prediction to these advanced applications. We also suggest a set of guidelines for reporting AlphaFold predictions.

D.9.8 Protein language models and structure

- Weissenow and Rost (2025) — Discusses how embeddings from protein language models are replacing multiple sequence alignments as input for many prediction tasks, and the trade-offs in accuracy and cost. Directly relevant to the comparison of language models with alignments. *(review; CC BY-NC-ND 4.0 (link only))*

  Are protein language models the new universal key?. Weissenow & Rost (2025). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2025.102997

  > Protein language models (pLMs) capture some aspects of the grammar of the language of life as written in protein sequences. The so-called pLM embeddings implicitly contain this information. Therefore, embeddings can serve as the exclusive input into downstream supervised methods for protein prediction. Over the last 33 years, evolutionary information extracted through simple averaging for specific protein families from multiple sequence alignments (MSAs) has been the most successful universal key to the success of protein prediction. For many applications, MSA-free pLM-based predictions now have become significantly more accurate. The reason for this is often a combination of two aspects. Firstly, embeddings condense the grammar so efficiently that downstream prediction methods succeed with small models, i.e., they need few free parameters in particular in the era of exploding deep neural networks. Secondly, pLM-based methods provide protein-specific solutions. As additional benefit, once the pLM pre-training is complete, pLM-based solutions tend to consume much fewer resources than MSA-based solutions. In

> fact, we appeal to the community to rather optimize foundation models than to retrain new ones and to evolve incentives for solutions that require fewer resources even at some loss in accuracy. Although pLMs have not, yet, succeeded to entirely replace the body of solutions developed over three decades, they clearly are rapidly advancing as the universal key for protein prediction.

- Zhang et al. (2024) — A research article rather than a review: it analyses what ESM-2 has learned and argues that it mainly stores coevolution statistics of interacting motifs. Helps to understand why language-model structure prediction resembles alignment-based methods and where it fails. *(research article; CC BY 4.0)*

  Protein language models learn evolutionary statistics of interacting sequence motifs. Zhang et al. (2024). *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.2406285121

  > Protein language models (pLMs) have emerged as potent tools for predicting and designing protein structure and function, and the degree to which these models fundamentally understand the inherent biophysics of protein structure stands as an open question. Motivated by a finding that pLM-based structure predictors erroneously predict nonphysical structures for protein isoforms, we investigated the nature of sequence context needed for contact predictions in the pLM Evolutionary Scale Modeling (ESM-2). We demonstrate by use of a "categorical Jacobian" calculation that ESM-2 stores statistics of coevolving residues, analogously to simpler modeling approaches like Markov Random Fields and Multivariate Gaussian models. We further investigated how ESM-2 "stores" information needed to predict contacts by comparing sequence masking strategies, and found that providing local windows of sequence information allowed ESM-2 to best recover predicted contacts. This suggests that pLMs predict contacts by storing motifs of pairwise contacts. Our investigation highlights the limitations of current pLMs and underscores the importance of understanding the underlying mechanisms of these models.

D.9.9 Membrane proteins

- Phillips and Miller (2021) — A short review of how α-helical membrane proteins are inserted and folded in the endoplasmic reticulum and how misfolded ones are recognised. Gives the biological context for topology and insertion. *(review; CC BY 4.0)*

  Membrane protein folding and quality control. Phillips & Miller (2021). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2021.03.003

  > Membrane proteins account for a quarter of cellular proteins, and most are synthesised at the endoplasmic reticulum (ER). Insertion and folding of polypeptides in the membrane environment is prone to error, necessitating diverse quality control systems. Recent discoveries have demonstrated how forces act on the nascent chain during insertion, and revealed new translocon components and accessories that facilitate the correct biogenesis of substrates. Our understanding of one of the best studied quality control systems-ER-associated degradation-has been advanced through new structural and functional studies of the core Hrd1 complex, and through the discovery of a new branch of this degradative pathway. New data also reveal how cells resolve clogged translocons, which would otherwise be unable to function. Finally, new work elucidates how mitochondrial tail-anchored proteins that have been mistargeted to the ER are identified and destroyed. Overall, we describe an emerging picture of an increasingly complex quality control network.

- Whitley et al. (2021) — Explains how the translocon recognises transmembrane helices, the hydrophobicity-based insertion scale and the role of charged residues in topology. Underpins the principles behind α-helical topology prediction. *(review; CC BY 4.0)*

  Folding and Insertion of Transmembrane Helices at the ER. Whitley et al. (2021). *International Journal of Molecular Sciences*. doi:10.3390/ijms222312778

  > In eukaryotic cells, the endoplasmic reticulum (ER) is the entry point for newly synthesized proteins that are subsequently distributed to organelles of the endomembrane system. Some of these proteins are completely translocated into the lumen of the ER while others integrate stretches of amino acids into the greasy 30 Å wide interior of the ER membrane bilayer. It is generally accepted that to exist in this non-aqueous environment the majority of membrane integrated amino acids are primarily non-polar/hydrophobic and adopt an α-helical conformation. These stretches are typically around 20 amino acids long

> and are known as transmembrane (TM) helices. In this review, we will consider how transmembrane helices achieve membrane integration. We will address questions such as: Where do the stretches of amino acids fold into a helical conformation? What is/are the route/routes that these stretches take from synthesis at the ribosome to integration through the ER translocon? How do these stretches 'know' to integrate and in which orientation? How do marginally hydrophobic stretches of amino acids integrate and survive as transmembrane helices?

- Horne et al. (2020) — Reviews the structure, folding and membrane insertion of β-barrel outer membrane proteins and the role of the BAM complex and the lipid bilayer. Covers the β-barrel subsection. *(review; CC BY 4.0)*

  Role of the lipid bilayer in outer membrane protein folding in Gram-negative bacteria. Horne et al. (2020). *Journal of Biological Chemistry*. doi:10.1074/jbc.rev120.011473

  > β-Barrel outer membrane proteins (OMPs) represent the major proteinaceous component of the outer membrane (OM) of Gram-negative bacteria. These proteins perform key roles in cell structure and morphology, nutrient acquisition, colonization and invasion, and protection against external toxic threats such as antibiotics. To become functional, OMPs must fold and insert into a crowded and asymmetric OM that lacks much freely accessible lipid. This feat is accomplished in the absence of an external energy source and is thought to be driven by the high thermodynamic stability of folded OMPs in the OM. With such a stable fold, the challenge that bacteria face in assembling OMPs into the OM is how to overcome the initial energy barrier of membrane insertion. In this review, we highlight the roles of the lipid environment and the OM in modulating the OMP-folding landscape and discuss the factors that guide folding in vitro and in vivo We particularly focus on the composition, architecture, and physical properties of the OM and how an understanding of the folding properties of OMPs in vitro can help explain the challenges they encounter during folding in vivo Current models of OMP biogenesis in the cellular environment are still in flux, but the stakes for improving the accuracy of these models are high. OMP folding is an essential process in all Gram-negative bacteria, and considering the looming crisis of widespread microbial drug resistance it is an attractive target. To bring down this vital OMP-supported barrier to antibiotics, we must first understand how bacterial cells build it.

D.9.10 Dynamics and docking

- Braun et al. (2018) — Introduces the foundations of molecular simulation: force fields, integration, ensembles, sampling and statistical uncertainty. Written for beginners and useful for understanding what molecular dynamics can and cannot tell you. *(tutorial; CC BY 4.0)*

  Best Practices for Foundations in Molecular Simulations [Article v1.0]. Braun et al. (2018). *Living Journal of Computational Molecular Science*. doi:10.33011/livecoms.1.1.5957

  > This document provides a starting point for approaching molecular simulations, guiding beginning practitioners to what issues they need to know about before and while starting their first simulations, and why those issues are so critical. This document makes no claims to provide an adequate introduction to the subject on its own. Instead, our goal is to help people know what issues are critical before beginning, and to provide references to good resources on those topics. We also provide a checklist of key issues to consider before and while setting up molecular simulations which may serve as a foundation for other best practices documents.

- Togashi and Flechsig (2018) — Explains elastic network models and normal-mode analysis and how they are used to describe large-scale protein motions. Covers the subsection on normal modes. *(review; CC BY 4.0)*

  Coarse-Grained Protein Dynamics Studies Using Elastic Network Models. Togashi & Flechsig (2018). *International Journal of Molecular Sciences*. doi:10.3390/ijms19123899

  > Elastic networks have been used as simple models of proteins to study their slow structural dynamics. They consist of point-like particles connected by linear Hookean springs and hence are convenient for linear normal mode analysis around a given reference structure. Furthermore, dynamic simulations using these models can provide new insights. As the computational cost associated with these models is considerably lower compared to that of all-atom models, they are also convenient for comparative

studies between multiple protein structures. In this review, we introduce examples of coarse-grained molecular dynamics studies using elastic network models and their derivatives, focusing on the nonlinear phenomena, and discuss their applicability to large-scale macromolecular assemblies.

- Paggi et al. (2024) — A clear review of how docking works (sampling and scoring), why it succeeds or fails, and how deep-learning methods compare with physics-based docking. Directly relevant to the subsection on docking and scoring functions. *(review; CC BY 4.0)*

  The Art and Science of Molecular Docking. Paggi et al. (2024). *Annual Review of Biochemistry*. doi:10.1146/annurev-biochem-030222-120000

  > Molecular docking has become an essential part of a structural biologist's and medicinal chemist's toolkits. Given a chemical compound and the three-dimensional structure of a molecular target-for example, a protein-docking methods fit the compound into the target, predicting the compound's bound structure and binding energy. Docking can be used to discover novel ligands for a target by screening large virtual compound libraries. Docking can also provide a useful starting point for structure-based ligand optimization or for investigating a ligand's mechanism of action. Advances in computational methods, including both physics-based and machine learning approaches, as well as in complementary experimental techniques, are making docking an even more powerful tool. We review how docking works and how it can drive drug discovery and biological research. We also describe its current limitations and ongoing efforts to overcome them.

D.9.11 Protein design

- Pan and Kortemme (2021) — Explains the principles of physics-based de novo design, including energy functions and sequence and backbone design, and the early deep-learning approaches. Good for the physics-based design subsection. *(review; CC BY 4.0)*

  Recent advances in de novo protein design: Principles, methods, and applications. Pan & Kortemme (2021). *Journal of Biological Chemistry*. doi:10.1016/j.jbc.2021.100558

  > The computational de novo protein design is increasingly applied to address a number of key challenges in biomedicine and biological engineering. Successes in expanding applications are driven by advances in design principles and methods over several decades. Here, we review recent innovations in major aspects of the de novo protein design and include how these advances were informed by principles of protein architecture and interactions derived from the wealth of structures in the Protein Data Bank. We describe developments in de novo generation of designable backbone structures, optimization of sequences, design scoring functions, and the design of the function. The advances not only highlight design goals reachable now but also point to the challenges and opportunities for the future of the field.

- Winnifrith et al. (2024) — A short overview of generative models for protein design, including diffusion models for backbones, inverse folding and language models for sequences. Covers the deep-learning design subsection. *(review; CC BY 4.0)*

  Generative artificial intelligence for de novo protein design. Winnifrith et al. (2024). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2024.102794

  > Engineering new molecules with desirable functions and properties has the potential to extend our ability to engineer proteins beyond what nature has so far evolved. Advances in the so-called 'de novo' design problem have recently been brought forward by developments in artificial intelligence. Generative architectures, such as language models and diffusion processes, seem adept at generating novel, yet realistic proteins that display desirable properties and perform specified functions. State-of-the-art design protocols now achieve experimental success rates nearing 20%, thus widening the access to de novo designed proteins. Despite extensive progress, there are clear field-wide challenges, for example, in determining the best in silico metrics to prioritise designs for experimental testing, and in designing proteins that can undergo large conformational changes or be regulated by post-translational modifications. With an increase in the number of models being developed, this review provides a framework to understand how these tools fit into the overall process of de novo protein design. Throughout, we highlight the power of incorporating biochemical knowledge to improve performance and interpretability.

- Jänes and Beltrao (2024) — Reviews deep-learning structure prediction beyond single chains (complexes, alternative conformations, evolution) and its use in protein design. Links the design section back to the prediction methods it builds on. *(review; CC BY 4.0)*

  Deep learning for protein structure prediction and design—progress and applications. Jänes & Beltrao (2024). *Molecular Systems Biology*. doi:10.1038/s44320-024-00016-x

  > Proteins are the key molecular machines that orchestrate all biological processes of the cell. Most proteins fold into three-dimensional shapes that are critical for their function. Studying the 3D shape of proteins can inform us of the mechanisms that underlie biological processes in living cells and can have practical applications in the study of disease mutations or the discovery of novel drug treatments. Here, we review the progress made in sequence-based prediction of protein structures with a focus on applications that go beyond the prediction of single monomer structures. This includes the application of deep learning methods for the prediction of structures of protein complexes, different conformations, the evolution of protein structures and the application of these methods to protein design. These developments create new opportunities for research that will have impact across many areas of biomedical research.

The following open-access paper(s) are reproduced under their original licences.

frontiers | Frontiers in Bioinformatics

TYPE Mini Review
PUBLISHED 28 February 2023
DOI 10.3389/fbinf.2023.1120370

Check for updates

OPEN ACCESS

EDITED BY
Petras Kundrotas,
University of Kansas, United States

REVIEWED BY
Andrzej Kolinski,
University of Warsaw, Poland
Arun Prasad Pandurangan,
Department of Biochemistry, University of Cambridge, United Kingdom
Lim Heo,
Michigan State University, United States

*CORRESPONDENCE
Samantha K. Teixeira,
samantha.teixeira@hc.fm.usp.br

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Protein Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 09 December 2022
ACCEPTED 17 February 2023
PUBLISHED 28 February 2023

CITATION
Bertoline LMF, Lima AN, Krieger JE and Teixeira SK (2023), Before and after AlphaFold2: An overview of protein structure prediction.
Front. Bioinform. 3:1120370.
doi: 10.3389/fbinf.2023.1120370

COPYRIGHT
© 2023 Bertoline, Lima, Krieger and Teixeira. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Before and after AlphaFold2: An overview of protein structure prediction

Letícia M. F. Bertoline†, Angélica N. Lima†, Jose E. Krieger and Samantha K. Teixeira*†

Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of São Paulo Medical School, São Paulo, Brazil

Three-dimensional protein structure is directly correlated with its function and its determination is critical to understanding biological processes and addressing human health and life science problems in general. Although new protein structures are experimentally obtained over time, there is still a large difference between the number of protein sequences placed in Uniprot and those with resolved tertiary structure. In this context, studies have emerged to predict protein structures by methods based on a template or free modeling. In the last years, different methods have been combined to overcome their individual limitations, until the emergence of AlphaFold2, which demonstrated that predicting protein structure with high accuracy at unprecedented scale is possible. Despite its current impact in the field, AlphaFold2 has limitations. Recently, new methods based on protein language models have promised to revolutionize the protein structural biology allowing the discovery of protein structure and function only from evolutionary patterns present on protein sequence. Even though these methods do not reach AlphaFold2 accuracy, they already covered some of its limitations, being able to predict with high accuracy more than 200 million proteins from metagenomic databases. In this mini-review, we provide an overview of the breakthroughs in protein structure prediction before and after AlphaFold2 emergence.

KEYWORDS

protein structure prediction, AlphaFold, template-based modeling, free modeling, protein language model

1 Introduction

Protein is the term proposed by the Swedish chemist Jacob Berzelius to compounds containing nitrogen and constituted by a combination of amino acids linked by a bond, called peptide bond (Fruton, 1985; Wisniak, 2000; Lehninger et al., 2005; Voet et al., 2014). They are responsible for housekeeping and specific functions essential to life, such as cell structural support, immune protection, enzymatic catalysis, cell signal transduction to transcription and translation regulation (Pearce and Zhang, 2021). The biological function of a protein depends on its tertiary/quaternary structures that derives ultimately from the folding of a polypeptide sequence(s) considering physical chemistry principles and the lowest free energy level, being the understanding of protein folding one of the most important goals in structural biology (Lehninger et al., 2005; Voet et al., 2014).

Experimentally, tertiary protein structures are resolved by X-ray crystallography, nuclear magnetic resonance and electron cryomicroscopy (cryo-EM). Nevertheless, those techniques

are complex, time consuming, expensive, and often the structure is not in its native form. Under these constrains, it is no surprise that the number of proteins with resolved tertiary structures is small (200,988 PDB entries) compared to the large number of proteins sequence known (229,580,745 entries on UniProtKB as of January 2023). This is an open challenge waiting for innovative ways to develop novel protein structure prediction approaches such as the ones using computationally models to predict the three dimensional protein structures starting from a polypeptide sequence (Dill et al., 2007; Nassar et al., 2021).

Several algorithms and web servers have been developed with the aim to improve the protein structure prediction. The relevance of these efforts are underscored by the different application that shall be affected including rational drug design, mutational studies and structural comparison, evolutionary analysis, folding studies. Here, we provide an overview of the methods developed to perform protein structure prediction to compare with AlphaFold2, which combines neural networks and homology modeling to generate models that may have experimental accuracy in a growing number of examples. We also discuss its applicability, limitations and efforts to improve the overall algorithm's performance.

2 Structure prediction methods

The prediction methods are usually divided into template-based modeling (TBM) and free modeling (FM), considering the use or not of templates (Gromiha et al., 2018; Bongirwar and Mokhade, 2022; Paiva et al., 2022), even though, recently, some TBM methods use energy-guided model refinement, and part of FM uses fragment-based sampling approaches, extracting information from Protein Data Bank (PDB) through machine learning. Within these two groups, the algorithms developed are usually classified into three different groups: *ab initio* (a FM methodology), threading/fold recognition (a TBM methodology), and homology (a TBM methodology) (Gromiha et al., 2018; Agnihotry et al., 2022). Despite controversial, this classification corresponds to the categories from the Critical Assessment of Structure Prediction (CASP), a biennial competition with the aim to establish the current state of the art in protein structure prediction, which contains three categories: TBM, FM, and an intermediate category, FM/TBM. In this contest, participants submit their models for proteins whose experimental structure has not been published yet (Kryshtafovych et al., 2019).

The *ab initio* approaches are based on the thermodynamics hypothesis that the native protein structure presents the lowest free energy possible (Hardin et al., 2002; Yuan et al., 2003; Gromiha et al., 2018; Agnihotry et al., 2022). The idea of these methods are to predict new folds considering physicochemical properties from the protein fold process, such as hydrogen bonding, contact potential energy, PDB-derived secondary structure propensities, and folding involving both bonded and non-bonded interactions. The *ab initio* method may or may not take into account motif identification in databases using small fragments. Some reviews even divide this category into two distinct methodologies, one dependent and another independent on database information. In this mini-review, we will use the term *ab initio* for methods that model protein structure without a template and that use the laws of thermodynamics as a basis, even though FM methods commonly exploit the information from known structures.

The main advantage of *ab initio* methods is the capacity to obtain novel and unknown protein folds (Dorn et al., 2014). Nevertheless, the complexity of the problem and the high number of conformational possibilities is computationally demanding, limiting the use for long protein sequences. Methods that use fragments may help to reduce the conformational space, but again avoid the prediction of new protein folds. One example of *ab initio* method that considers fragments is QUARK, a template-free novel program developed by Xu and Zhang (2012) in 2012. Briefly, this program breaks the protein sequence into fragments of 20 amino acids, searches for structure in a database, and then, using replica-exchange Monte Carlo simulations, unites the fragments taking into account the force field and energy terms, constituting a complete model.

The threading/fold recognition methodologies are based on the idea that structure is more conserved than the amino acid sequence and has a limited number of protein structure folds in nature (Rost et al., 1997; Dorn et al., 2014; Gromiha et al., 2018; Agnihotry et al., 2022). They consist in choosing the best 3D template of known foldings that fit well in the target sequence considering a scoring function built on pairwise potential, second structure comparison, as well as solvent properties. Thus, the target sequence is aligned with the structure model with the optimal scoring function, reorganizing the atoms of the target sequence in the aligned backbone. Finally, the affinity of the sequence with three-dimensional fold is verified followed by a manual verification. GenTHREADER is a program that uses threading techniques to evaluate the alignments, made using a sequence profile method, and then generates models that will be evaluated by a neural network to give a confidence measure (Jones, 1999).

The homology models derive from the fact that two amino acid sequences that are highly similar have similar structures (Dorn et al., 2014; Gromiha et al., 2018; Agnihotry et al., 2022; Sanjeevi et al., 2022). For this, the target sequence is aligned to a sequence in which the structure is known and an atomic model for the target protein is generated taking into account its similarities with the template backbone, followed by modeling loop regions and sidechains. The final step is to submit the model to energy minimization and evaluate it using the Ramachandran plot. Usually, homology methods achieve protein structures with higher accuracy than other methodologies. But, as other TBM methods, they are limited due to their inability to predict structures for new proteins, as they are dependent on templates. SWISS-MODEL is an automated system that uses homology modeling to predict a three dimensional structure of a protein (Guex et al., 2009; Kiefer et al., 2009). It was the first web server available to generate a 3D protein structure. This program integrates and automates all the processes involved in a homology modeling method, creating a fully automated workflow using a PERL based framework (Kiefer et al., 2009). The program presents an interface friendly for non-bioinformatician users and information, such as PFAM domain annotation and other tools from SWISS-MODEL (Kiefer et al., 2009).

Recently, new hybrid techniques have been published combining tools or improving known methods with artificial intelligence approaches. This was possible in part due to the

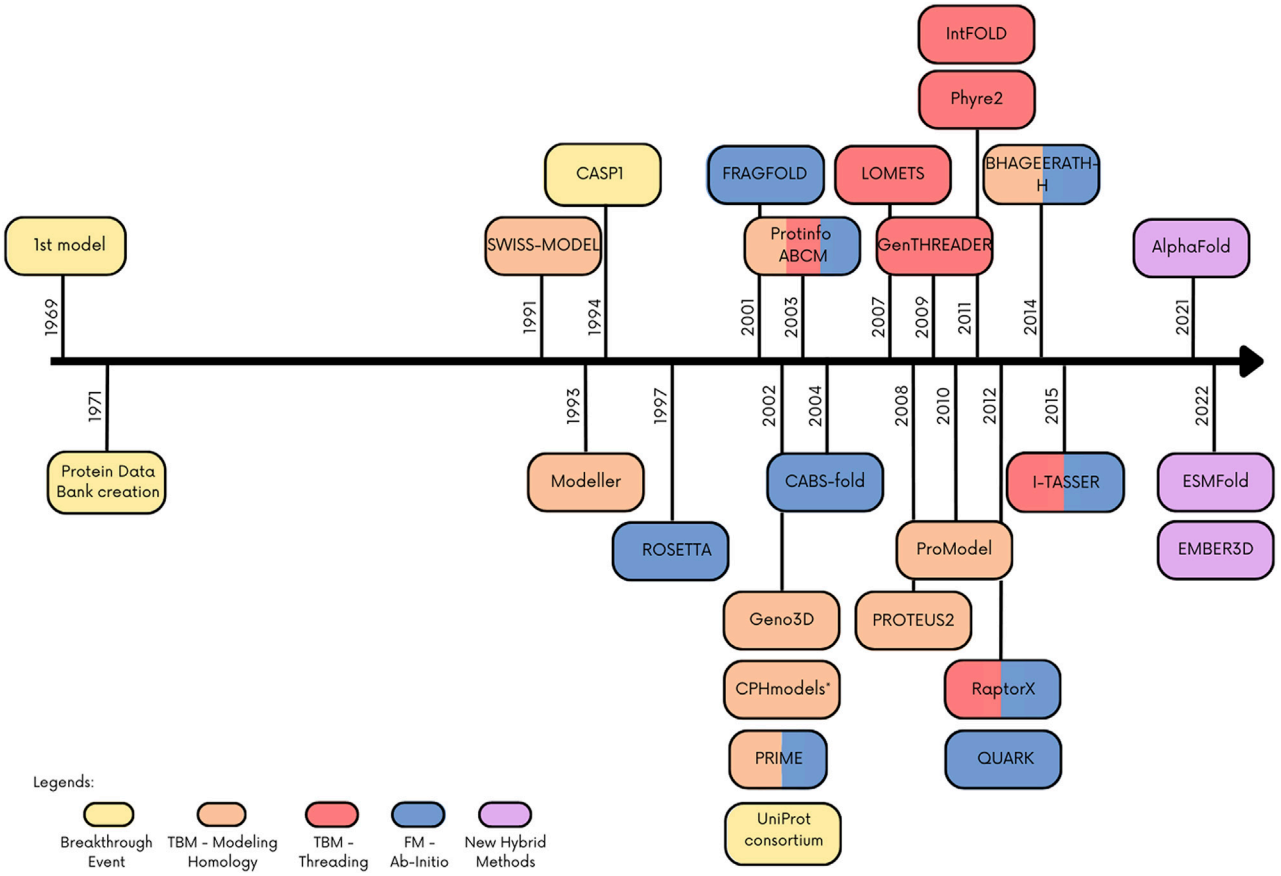


FIGURE 1
Timeline with main events and programs/webserver in the protein structure prediction. Colored boxes indicate the method or important event in the field.

development and improvement of computer processing. One of the breakthrough methods that not only combines methodology but also uses artificial intelligence is AlphaFold, an algorithm that beats the other tools in CASP13 and currently represents the state of art in protein structure prediction. Figure 1 depicts a timeline of the emergence of protein structure prediction programs/web servers and their classification considering the groups of methods adopted in this review, as well as other important dates for this field. More comprehensive and detailed reviews about protein structure prediction methods/tools can be found in Paiva et al. (2022), Bongirwar and Mokhade (2022).

3 AlphaFold

In 2018, DeepMind, a startup of Google, presented a new software that best performed in the 13th edition of CASP, named AlphaFold. In this competition, AlphaFold achieved the best position in the FM (best-of-five), reaching a summed z-score of 52.8 *versus* 36.6 from the second place and, combining FM and FM/TBM categories, achieved 68.3 z-score *versus* 48.2 (Senior et al., 2019). Even without using a template, AlphaFold also scored well in the TBM category (Senior et al., 2019).

The first version of AlphaFold used deep learning to predict the protein structure, demonstrating that it is possible to learn protein specific potential by training a neural network giving only the protein sequence. It contains a convolutional neural network that is trained by PDB structures to predict the distances between residues, creating distograms. From the amino acid sequence of the target protein, the neural network predicts a distogram based on multiple sequence alignment (MSA) features, in which a separate output from prediction network predicts the probability of backbone torsion distribution. The combined potential obtained by both ends is then optimized by gradient descent, predicting the protein structure itself (Senior et al., 2020).

Presented at CASP14 between May and July 2020, AlphaFold2 predicted protein structures with more accuracy than other competing methods, demonstrating a root-mean-square deviation (RMSD) among prediction and experimental backbone structures of 0.8Å *versus* the 2.8Å from the next best performing method. Moreover, AlphaFold2 scored 244.0 in summed z-scores compared with 90.8 for the next closest group (Jumper et al., 2021a; Jumper et al., 2021b). The great performance of AlphaFold2 in all Casp14 categories is depicted in Figure 2.

It is important to emphasize that AlphaFold2 contrasts considerably from the first version of AlphaFold. The authors

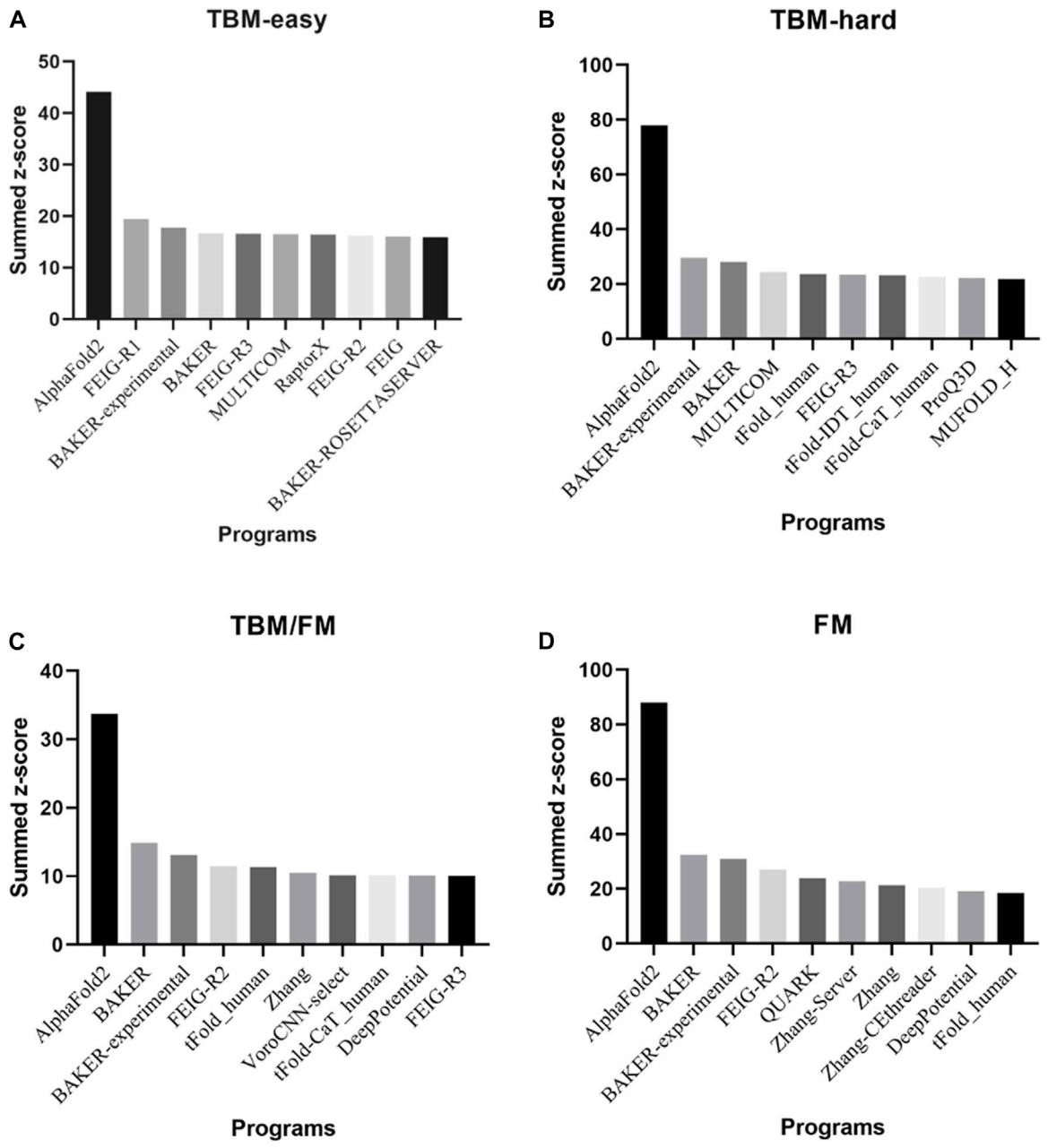


FIGURE 2
The top 10 programs and/or web servers in CASP14 in **(A)** TBM-easy, **(B)** TBM-hard, **(C)** TBM/FM, and **(D)** FM categories considering summed z-score. Data extracted by CASP official website.

attribute the high performance of AlphaFold2 by "incorporating novel neural networks architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures" (Jumper et al., 2021b; Bouatta et al., 2021; Callaway, 2022). AlphaFold2 uses as an input amino acid sequence to construct a MSA based on several databases of protein sequences to determine which parts of the sequence are mutation prone, detecting correlation between them. It also identifies proteins with similar structure with the input that will be used to build an initial representation of the target sequence (template), named as pair representation. Both strategies are not new and are shared by other algorithms in CASP14. Nevertheless, the breakthrough of AlphaFold2 is due to its neural network architectures, more specifically, the two neural network modules, evoformer and the structure module (Jumper et al., 2021b; Oxford Protein Informatics Group, 2021; Skolnick et al., 2021).

The evoformer extracts information from MSA and templates, exchanging information between them in flows back and forth throughout the network, improving the assessment of the MSA, that in turns modifies the protein structures hypothesized by the templates, allowing the MSA and templates in the correct "embedding space". It consists of two transformers, networks that

use attention to boost the speed with which a model can be trained, each of them specialized for a type of data, MSA or pair representations, with a clear communication channel between them. This allows the MSA transformer attention mechanism to incorporate information from the pair representation, adding a bias term from it, augmenting the attention mechanism and allowing it to pinpoint interacting pairs of residues. The pair representation transformer also works in a similar way, but includes an attention to terms of triangles of residues. The structure module, that also contains an attention architecture, uses both representations to prioritize the orientation of the protein backbone, considering the residue rotations and translations, localizing each side chain of each residue in highly constrained within a frame, followed by local refinement and minimization by gradient descent (Jumper et al., 2021b; Oxford Protein Informatics Group, 2021; Skolnick et al., 2021).

Protein-protein interactions are the basis of the biological process, and high-resolution structural characterization of these interactions give rise to insights of their molecular mechanisms and function, as well as direct the design of new drugs that are able to modulate these molecular pathways. Alphafold2 has been used to predict protein-protein interaction, using flexible linkers or artificial gaps and, in general, it predicted heterodimeric protein complexes accurately, exceeding docking approaches usually used in these analysis (Bryant et al., 2022; Yin et al., 2022). Nevertheless, it was limited to predict complexes, such as antigen-antibody and AlphaFold2 accuracy was also limited to predict complexes with protein from different species (Yin et al., 2022). In October 2021, DeepMind extended Alphafold2 to multiple chains - called AlphaFold-Multimer (Evans et al., 2022). For this, AlphaFold-Multimer was trained with protein complexes, and a series of changes in the code were made. The developers observed that performance was better in homomeric than heteromeric interfaces (Jumper et al., 2021b) and it did not predict binding of antigen to antibodies (Yin et al., 2022). Both, AlphaFold2 and AlphaFold-Multimer, are open-source codes and are available on github (https://github.com/deepmind/alphafold).

In the same year, in a partnership between DeepMind and the EMBL-European Bioinformatics Institute (EMBL-EBI), the AlphaFold Protein Structure Database (AlphaFold DB—Available in https://alphafold.ebi.ac.uk) was created, making available over 360,000 predicted structures from 21 organism proteomes (Varadi et al., 2022). Today, AlphaFold DB has over 200 million entries from the human and 47 other organism proteomes, with the structure predictions and their respective analyses freely available to the scientific community. Porta-Pardo et al. (2022) demonstrated that AlphaFold2 increases the structural coverage from 48% to 76% of all human protein residues, dropping the number of human protein without structural coverage from 5027 to 29. Moreover, they quantified that, among the 5027 of the proteins without structure previously, 4459 had structure prediction for over 50% of the protein's length (88,7%), in which 1408 with high-accuracy (28%). Despite the large amount of data, protein sequences containing non-standard amino acids, like selenocysteine, have been excluded, as well as multiple isoforms codifying by the same gene (Varadi et al., 2022). The database usability is easy, as input, protein name, gene, Uniprot accession number or organism can be used. As output, the AlphaFold DB provides the atomic coordinates in PDB and mmCIF formats and Predicted Aligned Error (PAEs) in JSON format. It is also possible to give feedback about the prediction structure through "Looks great" or "Could be improved" buttons.

Considered as the ground-breaking application of AI in science, AlphaFold has promised to revolutionize structure biology. Its application has been considered to design better protein expression experiment; To solve experimental structures faster, overcoming tedious model building, especially for X-ray crystallography, and to facilitate the interpretation of low-resolution cryo-EM; To protein design and drug development; To examine the effect of mutation in protein function, elucidating their potential impact on human diseases; To provide novel insights in poorly known molecular mechanisms (Perrakis and Sixma, 2021; Porta-Pardo et al., 2022). In this context, Noone et al. (2022) has demonstrated that indeed AlphaFold offers shortcuts to solve protein structures experimentally, predicting the remaining N-terminal region of PTX3 complex, which was, then, validated with cryo-EM class averaging. The hybrid cryoEM/AlphaFold structure allowed the mapping of the putative sites and regions of interaction, giving insights of the functions of PTX3.

However, despite its breakthrough accuracy and performance to predict protein structures, AlphaFold2 models have important limitations. First, AlphaFold2 has difficulty to predict intrinsically disordered proteins/regions (Ruff and Pappu, 2021) and loops (Stevens and He, 2022), especially considering the importance of the latter for drug screening and design, since they are exposed in protein surface and readily available to solvent and other proteins. Ruff and Pappu (2021) demonstrated that residues and regions predicted with low accuracy by AlphaFold2 overlaps intrinsically disordered regions, while Stevens and He (2022) showed that only short loops (<20 amino acids) are predicted with high accuracy by AlphaFold and it has the tendency to over-predict secondary structures in loop regions, usually alpha helix. Both regions are known to be hypervariable and flexible across orthologies, making it difficult to uncover evolutionary constraints from MSA.

Second, AlphaFold2 predicts only a single conformer, not identifying the apo and holo forms. In 67% of a dataset tested, AlphaFold prediction resembled holo form and the proteins were less predictable when the conformational differences between apo and holo forms increased (Saldaño et al., 2022). Moreover, Azzaz et al. (2022) demonstrated that structure prediction of membrane proteins by AlphaFold is not reliable, mainly because it presents inconsistencies in the location of the transmembrane domains. They stress that the protein environment influences the amino acid sequence, imposing folding constraints. These evidences together with the AlphaFold's inability to predict structures with metal ions, cofactors and other ligands, complexes with DNA or RNA, or post-translational modifications, such as glycosylation, methylation and phosphorylation (Perrakis and Sixma, 2021) highlight the steps to be overcome to improve AphaFold models in drug screening and design. Indeed, Scardino et al. (2023) demonstrated that AlphaFold models showed worse performance in high-throughput docking when compared to their corresponding experimental PDB structures, while Wong et al. (2022) showed that AlphaFold2 protein structure prediction exhibits weak performance on reverse docking in a search for binding targets of bacterial compounds, emphasizing that, even though AlphaFold2 provides rich structural information, more accurate

models of protein-ligand interactions are needed to improve use of AlphaFold2 for drug discovery.

Third, AlphaFold fails to predict defects in protein folding due to point mutations. As demonstrated by Buel and Walters (2022), the differences between mutated and wild-type models predicted by AlphaFold are very small, represented by backbones RMSD lower than 1Å. Moreover, Pak et al. (2021) also demonstrated that there is no correlation between AlphaFold accuracy metrics (pLDDT) and the impact of mutations on protein stability change (ΔΔG), neither with the side chain size change.

Finally, of AlphaFold2 cannot predict novel structures, since its algorithm is based on MSA and requires known structures databases. Another important aspect is that the use of evolutionary information from larger MSAs, requiring environmental systems and storage to detect the homology between known and target sequences, demands a powerful computing processors and its structure prediction is time consuming as protein length increases. To overcome this limitation Google offers the Google Colaboratory, which enables access to powerful GPUs. One of these solutions is ColabFold, a fast and easy-to-use software that replaces the AlphaFold2's homology by MMseq2 (Mirdita et al., 2022), making the computational demands less relevant.

4 New methods of protein structure prediction using protein language model

Recently, new free modeling methods have been published to overcome some of the limitations of AlphaFold2, such as the inability to predict novel structures and the necessity of high time and computing processes. These methods are based on protein language models, derived from natural language processing (NLP), which uses the amino acids sequence only and is able to learn evolutionary, structural and functional patterns derived from sequences available in databases, predicting a structural conformation. The idea behind those methods is that the amino acids correspond to words/tokens and proteins to sentences in NLP, assuming that similar semantics come from amino acids that occur in similar contexts.

There are three approaches used in language models: autoregressive, bidirectional, and masked. The first takes into count the previous tokens (amino acids) to predict the probability of a token, the second considers the previous and following tokens independently to estimate the probability of a token, and the last model considers all tokens in a sequence and replaces each token with a mask token. A synthesis of the recent advances in protein language modeling and their applications to protein prediction problems can be found elsewhere (Bepler and Berger, 2021).

Two methods have gained attention this year. ESMfold, developed by Lin et al. (2022) uses a masked transformer protein language model trained in deep information about biological properties, using 15 billion parameters. Compared with AlphaFold2, it did not present the same performance, achieving lower TM-scores (0.68 *versus* 0.85 using AlphaFold2 on CASP14). However, when evaluating AlphaFold2 without MSA, using only the amino acid sequence, ESMFold performed better (0.68 *versus* 0.37 using AlphaFold2 on Casp14). Moreover, it presented an accuracy comparable to AlphaFold2 for structures predicted with high confidence, demonstrating a median all-atom RMSD of 1.91Å and a backbone RMSD of 1.33Å, reaching similar experimental-level accuracy. Finally, this approach also demonstrated a significant improvement in prediction speed, since it does not require the construction of MSA. Using this approach, the authors presented the ESM Metagenomic Atlas, where they predicted more than 617 million structures from metagenomic databases, in which 225 million structures were predicted with high confidence, including those that are novel (Lin et al., 2022).

EMBER2 is a protein language-model developed by Weissenow et al. (2022a), that uses embeddings to predict inter-residue distance (2D structure) introducing attention heads derived from a pre-trained protein language model instead of MSA. Using EMBER2 in *trRosetta* to predict 3D structures is less accurate than AlphaFold2 in predicting a native structure and presents an inferior TM score (0.5 *versus* 0.79 in ColabFold) (Yang et al., 2020). However, it is faster than ColabFold by about 35 fold, similarly to ESMfold. As the comparison made in this work was not fair, since EMBER2 predicts 2D structures and AlphaFold, 3D structures, the authors developed a new approach that uses EMBER2 model, but now applied in three-dimensional structure prediction, named EMBER3D (Weissenow et al., 2022b). Again, Ember3D did not outperform AlphaFold, but is much faster than it and ESMFold. Whereas AlphaFold2 do not perform efficiently in the study of the impact of single amino acids variants into protein structure, Weissenow et al. (2022b) demonstrated that the differences in predicted distances maps generated by EMBER3D correlated well with native and mutant 3D structures from deep mutational scanning, having a better result than ESMfold. Furthermore, they developed a tool that presents the difference between native and mutant predicted structures by all possible amino acid exchanges in each position of a protein sequence. The similarity between de native amino acid and the mutated one helps the identification of exchanges that may cause a high impact on the protein structure.

These new approaches highlight the powerful capacity of language models to identify evolutionary, structural and functional patterns from massive protein sequence databases to solve protein prediction problem, improving prediction speed and requiring less computational power. It is expected that those approaches will develop and gain accuracy with the inclusion of biological knowledge and multi task learning.

5 Conclusion

Three-dimensional protein structure determination is important to elucidate the protein function, being critical to understanding biological processes and addressing human health and life science problems in general. Due to the difficulty of determining protein structures by experimental methods, their predictions have been one of the central problems of the scientific community. The advent of AlphaFold2 and the release of millions of protein structures predicted with high accuracy and available in AlphaFold DB allowed an unprecedented expansion of different research fields in life science, impacting the most biological

sciences, followed by biochemistry and cell biology, genetics, medical and health sciences and chemical sciences (Varadi and Velankar, 2022).

Protein structure prediction can be applied from understanding the interaction between pathogen and host, how pathogens survive and reproduce, and why they are resistant to certain drugs used, to the development of new and more efficient drugs, reducing the cost in drug discovery and development, as well as the development of new and improved vaccines (Duran-Frigola et al., 2013; Hazra and Patra, 2021). The development of a new vaccine strategy used against COVID-19 by Pfizer, Moderna and Johnson & Johnson, in which the prediction of the mutations and its potential effects in spike protein domain allowed the generation of an immunogen, highlights the window of opportunity that AlphaFold2 offers in the structure-guided vaccine design (Higgins, 2021).

The recent advances in protein structure prediction have contributed to improve protein folding issues. The emergence of AlphaFold2 has taken the problem of protein structure prediction to another level, reaching similar experimental-level accuracy in some cases. Nevertheless, improvements are needed to overcome the limitation to predict novel structures, intrinsically disordered regions and loops, the ability to predict only a single conformer without ligands and still present inconsistencies in its models, and the inability to predict the impact of missense mutation on protein structure. These limitations, ultimately, are important drawbacks to expand the use of AlphaFold in life science.

Efforts are in progress to improve AlphaFold performance and models. Johansson-Åkhe and Wallner (2022) demonstrated that randomly perturbing the neural network weights, forcing it to sample more conformational spaces can improve AlphaFold Multimer performance. Terwilliger et al. (2022) suggested that inclusion of new experimental information can improve parts of the models, showing that application of experimental density maps used iteratively allows the rebuilding of models that can be used as templates by AlphaFold new prediction. Finally, Hekkelman et al. (2022) enriched the models in the AlphaFold DB through transplantation of small molecules and ions based on homologous protein structures. They presented a new resource, the AlphaFill databank, to overcome the limitation presented by AlphaFold models that do not present ligands and co-factors, to help life scientists test new hypotheses and design target experiments.

Simultaneously, new strategies using protein language models are arising to compete with AlphaFold2 in terms of performance and accuracy and to overcome some of its limitations. These strategies, more specific ESMfold, offers an opportunity to identify new proteins and novel functions, allowing the identification of new species, including microorganisms and viruses that endanger human health, as well as those that offer solution to mitigate environmental problems, such as the degradation of polluting and the development of transgenic microorganisms for more efficient product production. For this, ESM Metagenomic Atlas is an important 3D protein structures resource to be deeply investigated by the scientific community (Lin et al., 2022). Finally, the efforts to better predict how mutations affect protein structure using these new approaches are essential to gain insights in human genetic disease and further improve disease management and prediction. Today, these methods do not outperform AlphaFold2, but with the improvement of deep language models and the enrichment of these models with biological information through multi-task learning, they promise to revolutionize the protein structural biology.

Author contributions

LB: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content. AL: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content. JK: Revising it critically for important intellectual content, and final approval of the version to be published ST: Conception and design, literature revision, drafting the manuscript and revising it critically for important intellectual content, and final approval of the version to be published.

Funding

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [INCT—2014/50889-7 to JK], the Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq [INCT—465586/2014-7 to JK], the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES [PROEX—169192/2018-0 to LB] and the Zerbini Foundation and Foxconn Brazil as part of a research grant “Machine Learning in Cardiovascular Medicine”.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agnihotry, S., Pathak, R. K., Singh, D. B., Tiwari, A., and Hussain, I. (2022). “Protein structure prediction,” in *Bioinformatics* (Amsterdam, Netherlands: Elsevier), 177–188. doi:10.1016/B978-0-323-89775-4.00023-7

Azzaz, F., Yahi, N., Chahinian, H., and Fantini, J. (2022). The epigenetic dimension of protein structure is an intrinsic weakness of the AlphaFold program. *Biomolecules* 12, 1527. doi:10.3390/biom12101527

Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell. Syst.* 12, 654–669.e3. doi:10.1016/j.cels.2021.05.017

Bongirwar, V., and Mokhade, A. S. (2022). Different methods, techniques and their limitations in protein structure prediction: A review. *Prog. Biophysics Mol. Biol.* 173, 72–82. doi:10.1016/j.pbiomolbio.2022.05.002

Bouatta, N., Sorger, P., and AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallogr. Sect. D. Struct. Biol.* 77, 982–991. doi:10.1107/S2059798321007531

Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* 13, 1265. doi:10.1038/s41467-022-28865-w

Buel, G., and Walters, K. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29, 1–2. doi:10.1038/s41594-021-00714-2

Callaway, E. (2022). What's next for AlphaFold and the AI protein-folding revolution. *Nature* 604, 234–238. doi:10.1038/d41586-022-00997-5

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007). The protein folding problem: When will it be solved? *Curr. Opin. Struct. Biol.* 17, 342–346. doi:10.1016/j.sbi.2007.06.001

Dorn, M., E Silva, M. B., Buriol, L. S., and Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Comput. Biol. Chem.* 53, 251–276. doi:10.1016/j.compbiolchem.2014.10.001

Duran-Frigola, M., Mosca, R., and Aloy, P. (2013). Structural systems pharmacology: The role of 3D structures in next-generation drug development. *Chem. Biol.* 20, 674–684. doi:10.1016/j.chembiol.2013.03.004

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2.

Fruton, J. S. (1985). Contrasts in scientific style. Emil fischer and franz hofmeister: Their research groups and their theory of protein structure. *Proc. Am. Philos. Soc.* 129, 313–370.

Gromiha, M. M., Nagarajan, R., and Selvaraj, S. (2018). Protein structural bioinformatics: An overview. *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.* 1–3, 445–459. doi:10.1016/B978-0-12-809633-8.20278-1

Guex, N., Peitsch, M. C., and Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-model and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30, S162–S173. doi:10.1002/elps.200900140

Hardin, C., Pogorelov, T. V., and Luthey-Schulten, Z. (2002). *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol.* 12, 176–181. doi:10.1016/S0959-440X(02)00306-8

Hazra, S., and Patra, S. (2021). *Importance of protein structure and function in pathogenesis: Highlights on the multifaceted organism* Mycobacterium tuberculosis. Amsterdam, Netherlands: Elsevier. doi:10.1016/b978-0-12-820084-1.00030-2

Hekkelman, M., de Vries, I., Joosten, R., and Perrakis, A. (2022). AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20, 205–213. doi:10.1038/s41592-022-01685-y

Higgins, M. K. (2021). Can we AlphaFold our way out of the next pandemic? *J. Mol. Biol.* 433, 167093. doi:10.1016/j.jmb.2021.167093

Johansson-Åkhe, I., and Wallner, B. (2022). Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front. Bioinforma.* 2, 959160. doi:10.3389/fbinf.2022.959160

Jones, D. T. (1999). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815. doi:10.1006/jmbi.1999.2583

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021a). Applying and improving AlphaFold at CASP14. *Proteins Struct. Funct. Bioinforma.* 89, 1711–1721. doi:10.1002/prot.26257

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021b). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 37, 387–392. doi:10.1093/nar/gkn750

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—round XIII. *Proteins Struct. Funct. Bioinforma.* 87, 1011–1020. doi:10.1002/prot.25823

Lehninger, A., Nelson, D., and Cox, M. (2005). *Principles of biochemistry*. 4th editio. New York.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v2.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi:10.1038/s41592-022-01488-1

Nassar, R., Dignon, G. L., Razban, R. M., and Dill, K. A. (2021). The protein folding problem: The role of theory. *J. Mol. Biol.* 433, 167126. doi:10.1016/j.jmb.2021.167126

Noone, D. P., Dijkstra, D. J., van der Klugt, T. T., van Veelen, P. A., de Ru, A. H., Hensbergen, P. J., et al. (2022). PTX3 structure determination using a hybrid cryoelectron microscopy and AlphaFold approach offers insights into ligand binding and complement activation. *PNAS* 33, e2208144119. doi:10.1073/pnas.2208144119

Oxford Protein Informatics Group (2021). *AlphaFold 2 is here: what's behind the structure prediction miracle*. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction- miracle/.

Paiva, V. A., Gomes, I. S., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., et al. (2022). Protein structural bioinformatics: An overview. *Comput. Biol. Med.* 147, 105695. doi:10.1016/j.compbiomed.2022.105695

Pak, M., Markhieva, K., Novikova, M., Petrov, D., Vorobyev, I., Maksimova, E., et al. (2021). *Using AlphaFold to predict the impact of single mutations on protein stability and function*. bioRxiv. doi:10.1101/2021.09.19.460937v1

Parto-Pardo, E., Ruiz-Serra, V., Valentini, S., and Valencia, A. (2022). The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.* 18, 1–17. doi:10.1371/journal.pcbi.1009818

Pearce, R., and Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* 297 (1), 100870. doi:10.1016/j.jbc.2021.100870

Perrakis, A., and Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep.* 22, e54046–6. doi:10.15252/embr.202154046

Rost, B., Schneider, R., and Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270, 471–480. doi:10.1006/jmbi.1997.1101

Ruff, K. M., and Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* 433, 167208. doi:10.1016/j.jmb.2021.167208

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., et al. (2022). Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* 38, 2742–2748. doi:10.1093/bioinformatics/btac202

Sanjeevi, M., Hebbar, P. N., Aiswarya, N., Rashmi, S., Rahul, C. N., Mohan, A., et al. (2022). *Methods and applications of machine learning in structure-based drug discovery*. Amsterdam, Netherlands: Elsevier. doi:10.1016/B978-0-323-90264-9.00025-8

Scardino, V., Di Filippo, J. I., and Cavasotto, C. N. (2023). How good are AlphaFold models for docking-based virtual screening? *iScience* 26, 105920. doi:10.1016/j.isci.2022.105920

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2019). Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* 87, 1141–1148. doi:10.1002/prot.25834

Skolnick, J., Gao, M., Zhou, H., and Singh, S. (2021). AlphaFold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J. Chem. Inf. Model.* 61, 4827–4831. doi:10.1021/acs.jcim.1c01114

Stevens, A. O., and He, Y. (2022). Benchmarking the accuracy of AlphaFold 2 in loop structure prediction. *Biomolecules* 12, 985. doi:10.3390/biom12070985

Terwilliger, T., Poon, B., Afonine, P., Schlicksup, C., Croll, T., Millan, C., et al. (2022). Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* 19, 1376–1382. doi:10.1038/s41592-022-01645-6

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50, D439–D444. doi:10.1093/nar/gkab1061

Varadi, M., and Velankar, S. (2022). The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, e2200128. doi:10.1002/pmic.202200128

Voet, D., Voet, J. G., and Charlotte, W. P. (2014). *Fundamental of biochemistry: Life at the molecular level*. 4th editio. Porto Alegre: wiley.

Weissenow, K., Heinzinger, M., and Rost, B. (2022a). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 30, 1169–1177.e4. doi:10.1016/j.str.2022.05.001

Weissenow, K., Heinzinger, M., Steinegger, M., and Rost, B. (2022b). Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. 1–16. doi:10.1101/2022.11.14.516473

Wisniak, J. (2000). Jons Jacob Berzelius A guide to the perplexed chemist. *Chem. Educ.* 5, 343–350. doi:10.1007/s00897000430a

Wong, F., Krishnan, A., Zheng, E., Stark, H., Manson, A., Earl, A. M., et al. (2022). Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* 18, e11081. doi:10.15252/msb.202211081

Xu, D., and Zhang, Y. (2012). *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma.* 80, 1715–1735. doi:10.1002/prot.24065

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1496–1503. doi:10.1073/pnas.1914677117

Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31, e4379. doi:10.1002/pro.4379

Yuan, X., Shao, Y., and Bystroff, C. (2003). *Ab initio* protein structure prediction using pathway models. *Comp. Funct. Genomics* 4, 397–401. doi:10.1002/cfg.305

D.10 Protein function

Further explanations for Chapter 10.

D.10.1 Definitions of function

- Thomas (2016) — Discusses what 'function' means for a gene product and how the GO distinguishes molecular function, biological process and cellular component. Useful background for the section's opening question of how to define function. *(book chapter; CC BY 4.0)*

  The Gene Ontology and the Meaning of Biological Function. Thomas (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_2

  > The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms. It is specifically designed for supporting the computational representation of biological systems. A GO annotation is an association between a specific gene product and a GO concept, together making a statement pertinent to the function of that gene. However, the meaning of the term "function" is not as straightforward as it might seem, and has been discussed at length in both philosophical and biological circles. Here, I first review these discussions. I then present an explicit formulation of the biological model that underlies the GO and annotations, and discuss how this model relates to the broader debates on the meaning of biological function.

- Gaudet et al. (2016) — A short introduction to the structure of the GO, its relations, annotations and evidence codes. A practical companion to the subsection on the Gene Ontology. *(primer; CC BY 4.0)*

  Primer on the Gene Ontology. Gaudet et al. (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_3

  > The Gene Ontology (GO) project is the largest resource for cataloguing gene function. The combination of solid conceptual underpinnings and a practical set of features have made the GO a widely adopted resource in the research community and an essential resource for data analysis. In this chapter, we provide a concise primer for all users of the GO. We briefly introduce the structure of the ontology and explain how to interpret annotations associated with the GO.

D.10.2 Function prediction

- Vu and Jung (2021) — Reviews GO-based function prediction from homology, domain and network methods to deep learning, including evaluation. Covers most of the subsections of this section in one place. *(review; CC BY 4.0)*

  Protein function prediction with gene ontology: from traditional to deep learning models. Vu & Jung (2021). *PeerJ*. doi:10.7717/peerj.12019

  > Protein function prediction is a crucial part of genome annotation. Prediction methods have recently witnessed rapid development, owing to the emergence of high-throughput sequencing technologies. Among the available databases for identifying protein function terms, Gene Ontology (GO) is an important resource that describes the functional properties of proteins. Researchers are employing various approaches to efficiently predict the GO terms. Meanwhile, deep learning, a fast-evolving discipline in data-driven approach, exhibits impressive potential with respect to assigning GO terms to amino acid sequences. Herein, we reviewed the currently available computational GO annotation methods for proteins, ranging from conventional to deep learning approach. Further, we selected some suitable predictors from among the reviewed tools and conducted a mini comparison of their performance using a worldwide challenge dataset. Finally, we discussed the remaining major challenges in the field, and emphasized the future directions for protein function prediction with GO.

- Cozzetto and Jones (2016) — Explains annotation transfer by sequence similarity, orthology and domain/family signatures, and the errors this can propagate. Useful for the subsections on homology- and family-based annotation. *(book chapter; CC BY 4.0)*

  Computational Methods for Annotation Transfers from Sequence. Cozzetto & Jones (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_5

  > Surveys of public sequence resources show that experimentally supported functional information is still completely missing for a considerable fraction of known proteins and is clearly incomplete for an even larger portion. Bioinformatics methods have long made use of very diverse data sources alone or in combination to predict protein function, with the understanding that different data types help elucidate complementary biological roles. This chapter focuses on methods accepting amino acid sequences as input and producing GO term assignments directly as outputs; the relevant biological and computational concepts are presented along with the advantages and limitations of individual approaches.

- Lin et al. (2024) — A recent survey grouping function predictors by the information they use (sequence, structure, networks, language models) and comparing them. Gives a more up-to-date view of the machine-learning methods. *(review; CC BY 4.0)*

  A comprehensive review and comparison of existing computational methods for protein function prediction. Lin et al. (2024). *Briefings in Bioinformatics*. doi:10.1093/bib/bbae289

  > Protein function prediction is critical for understanding the cellular physiological and biochemical processes, and it opens up new possibilities for advancements in fields such as disease research and drug discovery. During the past decades, with the exponential growth of protein sequence data, many computational methods for predicting protein function have been proposed. Therefore, a systematic review and comparison of these methods are necessary. In this study, we divide these methods into four different categories, including sequence-based methods, 3D structure-based methods, PPI network-based methods and hybrid information-based methods. Furthermore, their advantages and disadvantages are discussed, and then their performance is comprehensively evaluated and compared. Finally, we discuss the challenges and opportunities present in this field.

D.10.3 Subcellular localisation

- Imai and Nakai (2020) — Describes the main sorting signals (signal peptides, transit peptides, nuclear signals) and the tools that recognise them, written by authors of PSORT. Closely matches the section's structure from signals to prediction. *(review; CC BY 4.0)*

  Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences. Imai & Nakai (2020). *Frontiers in Genetics*. doi:10.3389/fgene.2020.607812

  > At the time of translation, nascent proteins are thought to be sorted into their final subcellular localization sites, based on the part of their amino acid sequences (i.e., sorting or targeting signals). Thus, it is interesting to computationally recognize these signals from the amino acid sequences of any given proteins and to predict their final subcellular localization with such information, supplemented with additional information (e.g., k -mer frequency). This field has a long history and many prediction tools have been released. Even in this era of proteomic atlas at the single-cell level, researchers continue to develop new algorithms, aiming at accessing the impact of disease-causing mutations/cell type-specific alternative splicing, for example. In this article, we overview the entire field and discuss its future direction.

- Gillani and Pollastri (2024) — A recent overview of localisation predictors, including deep-learning and language-model-based methods, and the datasets used to train them. *(review; CC BY 4.0)*

  Protein subcellular localization prediction tools <sup/>. Gillani & Pollastri (2024). *Computational and Structural Biotechnology Journal*. doi:10.1016/j.csbj.2024.04.032

> Protein subcellular localization prediction is of great significance in bioinformatics and biological research. Most of the proteins do not have experimentally determined localization information, computational prediction methods and tools have been acting as an active research area for more than two decades now. Knowledge of the subcellular location of a protein provides valuable information about its functionalities, the functioning of the cell, and other possible interactions with proteins. Fast, reliable, and accurate predictors provides platforms to harness the abundance of sequence data to predict subcellular locations accordingly. During the last decade, there has been a considerable amount of research effort aimed at developing subcellular localization predictors. This paper reviews recent subcellular localization prediction tools in the Eukaryotic, Prokaryotic, and Virus-based categories followed by a detailed analysis. Each predictor is discussed based on its main features, strengths, weaknesses, algorithms used, prediction techniques, and analysis. This review is supported by prediction tools taxonomies that highlight their rele- vant area and examples for uncomplicated categorization and ease of understandability. These taxonomies help users find suitable tools according to their needs. Furthermore, recent research gaps and challenges are discussed to cover areas that need the utmost attention. This survey provides an in-depth analysis of the most recent prediction tools to facilitate readers and can be considered a quick guide for researchers to identify and explore the recent literature advancements.

D.10.4 Predicting the effects of mutations

- Riccio et al. (2024) — A short, practical overview of variant effect predictors, how they are built and how to choose and interpret them. Relevant to the subsection on using predictions. *(review; CC BY 4.0)*

  Variant effect predictors: a systematic review and practical guide. Riccio et al. (2024). *Human Genetics*. doi:10.1007/s00439-024-02670-5

  > Large-scale association analyses using whole-genome sequence data have become feasible, but understanding the functional impacts of these associations remains challenging. Although many tools are available to predict the functional impacts of genetic variants, it is unclear which tool should be used in practice. This work provides a practical guide to assist in selecting appropriate tools for variant annotation. We conducted a MEDLINE search up to November 10, 2023, and included tools that are applicable to a broad range of phenotypes, can be used locally, and have been recently updated. Tools were categorized based on the types of variants they accept and the functional impacts they predict. Sequence Ontology terms were used for standardization. We identified 118 databases and software packages, encompassing 36 variant types and 161 functional impacts. Combining only three tools, namely SnpEff, FAVOR, and SparkINFERNO, allows predicting 99 (61%) distinct functional impacts. Thirty-seven tools predict 89 functional impacts that are not supported by any other tool, while 75 tools predict pathogenicity and can be used within the ACMG/AMP guidelines in a clinical context. We launched a website allowing researchers to select tools based on desired variants and impacts. In summary, more than 100 tools are already available to predict approximately 160 functional impacts. About 60% of the functional impacts can be predicted by the combination of three tools. Unexpectedly, recent tools do not predict more impacts than older ones. Future research should allow predicting the functionality of so far unsupported variant types, such as gene fusions.URL: https://cardio-care.shinyapps.io/VEP_Finder/ .Registration: OSF Registries on November 10, 2023, https://osf.io/s2gct .

- Livesey and Marsh (2023) — Benchmarks many predictors, including unsupervised evolutionary models, against deep mutational scanning data. A primary benchmark rather than a tutorial, but it illustrates the experimental-benchmark subsection well. *(review; CC BY 4.0)*

  Updated benchmarking of variant effect predictors using deep mutational scanning. Livesey & Marsh (2023). *Molecular Systems Biology*. doi:10.15252/msb.202211474

  > The assessment of variant effect predictor (VEP) performance is fraught with biases introduced by benchmarking against clinical observations. In this study, building on our previous work, we use independently generated measurements of protein function from deep mutational scanning (DMS) experiments for 26 human proteins to benchmark 55 different VEPs, while introducing minimal data circularity. Many top-performing VEPs are unsupervised methods including EVE, DeepSequence and ESM-1v, a protein language model that ranked first overall. However, the strong performance of recent supervised VEPs, in particular VARITY, shows that developers are taking data circularity and bias issues seriously. We also assess the performance of DMS and unsupervised VEPs for discriminating between known pathogenic and putatively benign missense variants. Our findings are mixed, demonstrating that some DMS datasets

perform exceptionally at variant classification, while others are poor. Notably, we observe a striking correlation between VEP agreement with DMS data and performance in identifying clinically relevant variants, strongly supporting the validity of our rankings and the utility of DMS for independent benchmarking.

- Gerasimavicius et al. (2025) — A brief review of how protein structure, including predicted structures, can be combined with sequence-based predictors to assess missense variants. *(review; CC BY 4.0)*

  Leveraging protein structural information to improve variant effect prediction. Gerasimavicius et al. (2025). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2025.103023

  > Despite massive sequencing efforts, understanding the difference between human pathogenic and benign variants remains a challenge. Computational variant effect predictors (VEPs) have emerged as essential tools for assessing the impact of genetic variants, although their performance varies. Initially, sequence-based methods dominated the field, but recent advances, particularly in protein structure prediction technologies like AlphaFold, have led to an increased utilization of structural information by VEPs aimed at scoring human missense variants. This review highlights the progress in integrating structural information into VEPs, showcasing novel models such as AlphaMissense, PrimateAI-3D, and CPT-1 that demonstrate improved variant evaluation. Structural data offers more interpretability, especially for non-loss-of-function variants, and provides insights into complex variant interactions in vivo. As the field advances, utilizing biomolecular complex structures will be pivotal for future VEP development, with recent breakthroughs in protein-ligand and protein-nucleic acid complex prediction offering new avenues.

D.10.5 Benchmarking function prediction

- Friedberg and Radivojac (2016) — Written by CAFA organisers; explains the design of CAFA, its metrics and the difficulties of evaluating function prediction with incomplete annotations. *(book chapter; CC BY 4.0)*

  Community-Wide Evaluation of Computational Function Prediction. Friedberg & Radivojac (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_10

  > A biological experiment is the most reliable way of assigning function to a protein. However, in the era of high-throughput sequencing, scientists are unable to carry out experiments to determine the function of every single gene product. Therefore, to gain insights into the activity of these molecules and guide experiments, we must rely on computational means to functionally annotate the majority of sequence data. To understand how well these algorithms perform, we have established a challenge involving a broad scientific community in which we evaluate different annotation methods according to their ability to predict the associations between previously unannotated protein sequences and Gene Ontology terms. Here we discuss the rationale, benefits, and issues associated with evaluating computational methods in an ongoing community-wide challenge.

- Škunca et al. (2016) — Discusses why assessing GO predictions is hard (open-world annotations, hierarchy, annotation biases) and how evaluation strategies address this. Matches the subsection on why function prediction is hard to evaluate. *(book chapter; CC BY 4.0)*

  Evaluating Computational Gene Ontology Annotations. Škunca et al. (2016). *The Gene Ontology Handbook*. doi:10.1007/978-1-4939-3743-1_8

  > Two avenues to understanding gene function are complementary and often overlapping: experimental work and computational prediction. While experimental annotation generally produces high-quality annotations, it is low throughput. Conversely, computational annotations have broad coverage, but the quality of annotations may be variable, and therefore evaluating the quality of computational annotations is a critical concern.In this chapter, we provide an overview of strategies to evaluate the quality of computational annotations. First, we discuss why evaluating quality in this setting is not trivial. We highlight the various issues that threaten to bias the evaluation of computational annotations, most of which stem from the incompleteness of biological databases. Second, we discuss solutions that address these issues, for example, targeted selection of new experimental annotations and leveraging the existing experimental annotations.

D.10.6 Experimental determination of function

- Bock et al. (2022) — A primer on pooled and arrayed CRISPR screens, their readouts and analysis; linked only. *(primer; not open access)*

  High-content CRISPR screening. Bock et al. (2022). *Nature Reviews Methods Primers*. doi:10.1038/s43586-021-00093-4

The following open-access paper(s) are reproduced under their original licences.

Protein function prediction with gene ontology: from traditional to deep learning models

Thi Thuy Duong Vu and Jaehee Jung

Department of Information and Communication Engineering, Myongji University, Yongin-si, Gyeonggi-do, South Korea

ABSTRACT

Protein function prediction is a crucial part of genome annotation. Prediction methods have recently witnessed rapid development, owing to the emergence of high-throughput sequencing technologies. Among the available databases for identifying protein function terms, Gene Ontology (GO) is an important resource that describes the functional properties of proteins. Researchers are employing various approaches to efficiently predict the GO terms. Meanwhile, deep learning, a fast-evolving discipline in data-driven approach, exhibits impressive potential with respect to assigning GO terms to amino acid sequences. Herein, we reviewed the currently available computational GO annotation methods for proteins, ranging from conventional to deep learning approach. Further, we selected some suitable predictors from among the reviewed tools and conducted a mini comparison of their performance using a worldwide challenge dataset. Finally, we discussed the remaining major challenges in the field, and emphasized the future directions for protein function prediction with GO.

Subjects Bioinformatics, Computational Biology, Data Mining and Machine Learning

Keywords Gene Ontology, Protein function prediction, Machine learning, Deep learning, CAFA3, Annotation

Submitted 13 April 2021
Accepted 29 July 2021
Published 24 August 2021

Corresponding author
Jaehee Jung, jhjung@mju.ac.kr

Academic editor
Vladimir Uversky

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.12019**

Copyright
2021 Vu and Jung

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Proteins are organic macromolecules that are fundamental determinants of the structure and function of living organisms. They play a role in numerous processes, including biochemical reactions, transmission of signals, nutrient transport, immune system boosting, etc. Therefore, understanding protein properties is essential not only from the biological and evolutionary perspectives, but also from the viewpoint of leveraging their potential in biomedical and pharmaceutical applications, and other areas (*Amiri-Dashatan et al., 2018*).

Generally, protein function identification is accomplished through manual or computational annotation. The former approach is the gold standard for functional annotation, because it is implemented by expert annotators and yields high quality curated results. Nonetheless, this approach is expensive and laborious, and thus, it is difficult to scale. Furthermore, because of the development of high-throughput sequencing technologies, such as next-generation sequencing (NGS), the amount of sequences to be annotated has increased dramatically. Thus, computational annotation methods have been developed as

How to cite this article Vu TTD, Jung J. 2021. Protein function prediction with gene ontology: from traditional to deep learning models. *PeerJ* **9**:e12019 http://doi.org/10.7717/peerj.12019

a must to automatically process the high volume of newly generated sequences, and also to improve the accuracy of the annotated data (*Sleator & Walsh, 2010*).

Because of the variability in the vocabulary used to define protein function, which makes the annotation process confusing to both humans and machines, various databases have been proposed to provide a standardized scheme, such as the Enzyme Commission (EC) (*Webb, 1992*), Functional Catalogue (FunCat) (*Ruepp et al., 2004*), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (*Kanehisa et al., 2004*). Currently, Gene Ontology (GO) (*Ashburner et al., 2000*; *Consortium, 2015*) is the most comprehensive resource, as it possesses all the desirable properties of a functional classification system (*Pandey, Kumar & Steinbach, 2006*). The GO consortium created a database for a controlled vocabulary describing the functional properties of genomic products (*e.g.*, genes, proteins, and RNA). Each ontology (vocabulary) belongs to one of three categories: molecular function (MF), biological process (BP), and cellular component (CC). In terms of structure, the GO follows a hierarchical organization as a directed acyclic graph (DAG), in which each term is a node and each edge connected to two nodes represents a parent–child relationship. It can be used to infer many types of information, such as "is-a" or "part-of". For instance, if term A is denoted as "is-a" of term B, it means that A is a sub-type of B. Further, "part-of" implies that the child node is necessarily part of the parent. This allows the flexible annotation of proteins with respect to the various levels of function—from general to specific terms—depending on the available evidence (*Stein, 2001*).

Automated function prediction (AFP) based on the GO system is a challenging problem in bioinformatics. Many studies have discussed protein functional annotation from different perspectives. Previous reviews focused on AFP (*Rost et al., 2003*) in terms of the data type used (*Watson, Laskowski & Thornton, 2005*; *Pandey, Kumar & Steinbach, 2006*; *Sleator & Walsh, 2010*; *Shehu, Barbará & Molloy, 2016*), drawbacks and corresponding solutions (*Friedberg, 2006*), protein interaction networks (*Sharan, Ulitsky & Shamir, 2007*), types of classified function (*Rentzsch & Orengo, 2009*), and GO assignment based on sequence information (*Cozzetto & Jones, 2017*). The two latest reviews are prepared by *Bonetta & Valentino (2020)* and *Zhao et al. (2020)*. *Bonetta & Valentino (2020)* demonstrate protein function prediction in the machine learning workflow. Meanwhile, *Zhao et al. (2020)* discuss prediction of gene function prediction from the GO modeling perspective. All these researches provided independent perspectives to the issue, but we observed that there has not been any detailed review about deep learning, which is an emerging approach for protein function prediction with GO terms. Therefore, our paper suggests the possibility of predicting protein function using both conventional learning and deep learning, further indicating that better predictive performance can be expected by comparing several methods with each other.

Herein, we reviewed automated protein function prediction using GO terms (ranging from traditional solutions to the most recently developed deep learning-based tools). We presented not only an overview of the literature, but also, a performance comparison of the emerging solutions. Further, we highlighted the challenges and future prospects of the field. We hope that this review will provide bioinformaticians, computer scientists

or any researchers who are interested in this topic, with the latest updates in terms of protein features used, models, and evaluation criteria, hoping to contribute to further improvements in the future.

SURVEY METHODOLOGY

Based on previous studies (*Jung & Thon, 2006*; *Jung et al., 2010*) and the aforementioned surveys, our review is divided in two main parts offering an overview of the field (from its inception to its present state). The first parts covers the conventional approach and includes solutions that do not use deep learning, while the second part describes methods that rely on deep learning to address the problem of protein functional annotation. We used Google Scholar https://scholar.google.com/ as the literature database to retrieve relevant publications, without applying any restrictions as to the publishing data, journal, or publisher.

Below, we review the conventional approach briefly, as some excellent literature reviews are already available on the subject (*Shehu, Barbará & Molloy, 2016*; *Cozzetto & Jones, 2017*). In that part, we summarize three main sub-categories of the traditional approach, and mention prominent or most recent corresponding studies. The second part of the review is our main focus. Search results using combinations of the following keywords: "deep learning", "deep neural network", "Gene Ontology", "GO terms", "protein function prediction", "functional annotation", and "gene function prediction", are selected after screening the titles, abstracts and conclusions. We also referred to references cited in the downloaded papers, to capture significant studies in the field.

CONVENTIONAL APPROACH FOR PREDICTING PROTEIN GO TERMS

In general, several techniques have been proposed to tackle the assignment of GO terms to proteins using different types of data and methods. We review the representative solutions below, focusing on three categories: similarity-based methods, probabilistic methods, and machine learning methods.

Similarity-based methods

Initially, functional annotations were assigned to uncharacterized proteins based on a simple principle: a protein sequence was searched in databases of experimentally curated proteins and if there were any similar proteins retrieved ("similar" in a specifically defined way), GO terms associated with the retrieved sequences were assigned to the query protein.

The very first solution of similarity-based methods was homology-based. These tools relied on local alignment search tools, such as the Basic Local Alignment Search Tool (BLAST) (*Altschul et al., 1997*). In these methods, the unknown sequence is searched in a database that curates well-annotated proteins. Then, a retrieved sequence with the highest alignment score—as per a predetermined threshold—is identified and its annotation is transferred to the query protein. OntoBlast (*Zehetner, 2003*), GOFigure (*Khan et al., 2003*), GOblet (*Hennig, Groth & Lehrach, 2003*), and GOtcha (*Martin, Berriman & Barton, 2004*)

are typical annotation systems adopting sequence similarity determined by BLAST search. Further details regarding these tools are presented in Table 1.

PFP (*Hawkins et al., 2009*) is another method that utilizes functional information associated with remote homologs by employing Position-Specific Iterated BLAST (PSI-BLAST) (*Altschul et al., 1997*). The tool is an improvement in terms of coverage and accuracy, as determined by analyzing a benchmark dataset, compared with GOtcha, top PSI-BLAST, and InterProScan (*Zdobnov & Apweiler, 2001*). Another variant solution based on sequence alignment is INGA (*Piovesan et al., 2015*), in which protein–protein interaction (PPI) network data is combined with domain assignment and sequence similarity from BLAST, to attain a consensus prediction of GO functions using enrichment analysis. Further, GoFDR (*Gong, Ning & Tian, 2016*) takes relevant GO terms from multiple sequence alignment (MSA) querying *via* BLAST or PSI-BLAST search. The probability of assigning a term to the query sequence is determined by the functionally discriminating residues (FDRs), a position-specific scoring matrix (PSSM) for the FDRs, and a score-to-probability table prepared using training sequences.

Although methods based on a local alignment search are straightforward and perform well to some extent, they also have some drawbacks, including database annotation errors or excessive function transfer, threshold relativity, and low sensitivity or specificity (*Sasson, Kaplan & Linial, 2006*). In addition to the alignment-based methods, several other predictors are available that transfer function annotations based on the similarity at the level of protein structure, protein family, or phylogenomics (*Shehu, Barbará & Molloy, 2016*).

Probabilistic methods

A number of probabilistic models have been devised for protein function deduction (*Deng, Chen & Sun, 2004a*; *Deng et al., 2004b*; *Letovsky & Kasif, 2003*; *Nariai, Kolaczyk & Kasif, 2007*). *Letovsky & Kasif (2003)* employed a functional linkage graph (*Marcotte et al., 1999*) constructed based on a PPI network of the yeast *Saccharomyces cerevisiae*. The working assumption was that the probability of sharing functions between (nodes) proteins in close proximity on the graph is higher than that for nodes that are not in close proximity. In this method, the probabilities of GO terms assigned to a protein sequence are derived from a binomial model incorporating the Markov Random Field (MRF) algorithm. Later, *Nariai, Kolaczyk & Kasif (2007)* integrated multiple sources of evidence (PPI, gene expression data, protein motif information, mutant phenotype data, and protein localization data) using a Bayesian framework, to improve the prediction performance, compared with their model that only uses PPI.

Working on the same model organism, *S. cerevisiae*, as used in the methods described above, *Kourmpetis et al. (2010)* suggested applying a Bayesian approach to the MRF model. This suggestion was considered an improvement for the estimation of model parameters and the provision of predictions from network data. Further, *Pinoli, Chicco & Masseroli (2015)* compared different weighting schemes as combined with three algorithms, namely, truncated singular value decomposition (tSVD), semantically improved tSVD (SIM), and probabilistic latent semantic analysis with normalization (pLSAnorm). The first two

Table 1 GO annotation predictors using the traditional approach.

Methods	Program name	Description	Reference
Similarity based	OntoBlast	Web server, which is a part of the "Ontologies TO GenomeMatrix" tool (*Hewelt et al., 2002*). It predicts protein function by a weighted list of GO terms associated with BLAST hit sequences in nine genome databases.	*Zehetner (2003)*
	GOFigure	Produces output as a clickable graph in four steps, including homologous sequence search, minimum cover graph construction, and assigning ontologies after scoring them.	*Khan et al. (2003)*
	GOblet	Software package allowing the user to define the *E*-value sensitivity and databases used for the BLAST search. All ontologies identified after matching the proteins are constructed in a summary DAG structure, where the number of sequences sharing a common GO term is cumulative to present the significance of the term.	*Hennig, Groth & Lehrach (2003)*
	Gotcha	Transfers GO associations retrieved from BLAST to gene products through a novel ranking scheme. The E-values of the derived terms are propagated to their parent terms on DAG, providing normalized confidence scores for an individual GO ontology.	*Martin, Berriman & Barton (2004)*
	PFP	Functional annotation system based on PSI-BLAST hits, exploiting matched sequences up to an unconventionally high threshold, and mining highly relevant GO terms from UniProt to score functions for unknown proteins.	*Hawkins et al. (2009)*
	INGA	Functional annotation server derives GO predictions from combining many data sources.	*Piovesan et al. (2015)*
	GoFDR	Multiple alignment-based method utilizing FDRs and PSSM to rank GO terms for amino acid sequences.	*Gong, Ning & Tian (2016)*
Probabilistic	–	Probabilistic framework based on PPI, predicting GO assignments for 10% unannotated proteins in yeast.	*Letovsky & Kasif (2003)*
	–	Leverages heterogeneous data to increase the probabilistic based model for the functional prediction of *Saccharomyces cerevisiae*.	*Nariai, Kolaczyk & Kasif (2007)*
	BMRF	Variant of the MRF-based approach on PPI that infers novel predictions for unannotated proteins.	*Kourmpetis et al. (2010)*
Machine learning	GOPET	GO term prediction and evaluation tool, supplying molecular function and biological process GO terms for any organism.	*Vinayagam et al. (2006)*
	PoGO	Ensemble model with InterPro sequence similarity, biochemical property, and protein tertiary structure to predict the GO of fungal proteins.	*Jung et al. (2010)*
	FFPred3	SVM-based tool, providing the homology-independent assignment of GO terms for eukaryotic proteins.	
	PANNZER2	Weighted *k*-nearest neighbor classifier, providing functional annotation for uncharacterized proteins with GO terms and free text descriptions.	*Törönen, Medlar & Holm (2018)*
	DeepText2GO	Combination of a text-based method and a sequence-based method to improve large-scale protein function prediction.	*You, Huang & Zhu (2018a)*
	NetGo	Effective LTR-based web server, combining both sequence and massive network information of proteins to annotate gene products.	*You et al. (2019)*

methods are based on linear algebra, while the latter is a modified probabilistic model based on Bayesian inference. These methods were successfully used to generate novel GO annotations for three model organisms, *i.e., Bos taurus*, *Danio rerio*, and *Drosophila melanogaster*.

Machine learning methods

Machine learning-based tools have been developed to identify the hidden relationships between various protein features (sequence, structure, or other related evolutionary evidence) and functional labels, based on a training set (a group of fully characterized macromolecules), and use that information to generate annotations for novel proteins. Various methods that rely on machine learning have been suggested, incorporating different lines of evidence as features to train the classifiers and predict GO terms. Branch (a) of Fig. 1 formulates the common workflow of those solutions, the details of which can vary between methods as described below.

Utilizing multiple support vector machine (SVM) classifiers is a predominant selection in several studies. For instance, GOPET (*Vinayagam et al., 2006*) feeds several features related to GO terms (sequence similarity measures based on BLAST search, frequency, quality of homolog annotation, and annotation level within the GO hierarchy) to 99 SVM classifiers. A specific term will be predicted to be a "correct" or "incorrect" label for an unknown sequence, with voting confidence scores calculated. As another example, FFPred (*Lobley et al., 2007*; *Lobley et al., 2008*) was initially established for the unannotated human proteome, but demonstrated a generalization to other proteomes. The latest version of FFPred is FFPred3 (*Cozzetto et al., 2016*), which is still based on SVM, is expanded to investigate the correlations between feature characteristics retrieved from sequences and structures within the three sub-ontologies (MF, BP, and CC). Another tool, Prediction of Gene Ontology terms (PoGO) (*Jung et al., 2010*), has been developed from Automatic Annotation of Protein Functional Class (AAPFC) (*Jung & Thon, 2006*). Instead of only using InterPro terms as features, PoGO integrates three more sources (sequence similarity, biochemical properties, and protein tertiary structure). Subsequently, SVM and a linear classifier (*Freund & Schapire, 1997*) are used as base-level classifiers before the meta-learning step.

Using the k-nearest neighbor (k-NN) algorithm, PANNZER2 (*Törönen, Medlar & Holm, 2018*) provides a fast functional annotation systems based on sequence homology and other annotation predictors. Meanwhile, MS-k NN (*Lan et al., 2013*) aggregates heterogeneous data to propose a competitive model for protein function prediction. DeepText2GO (*You, Huang & Zhu, 2018a*) is a consensus approach, integrating deep semantic text representation from MEDLINE citations (*ncb, 2018*) as text information, and sequence information obtained *via* BLAST and InterProScan. These features are fed to the k-NN and logistic regression models to produce the functions of no-knowledge proteins on a large scale. NetGO (*You et al., 2019*) is an extension of GoLabeler (*You et al., 2018b*), which employs the learning-to-rank(LTR) model to integrate sequence-based evidence. NetGO improves the performance of a large scale AFP by accessing the enormous protein–protein network of over 2000 species in the STRING database (*Szklarczyk et al., 2016*).

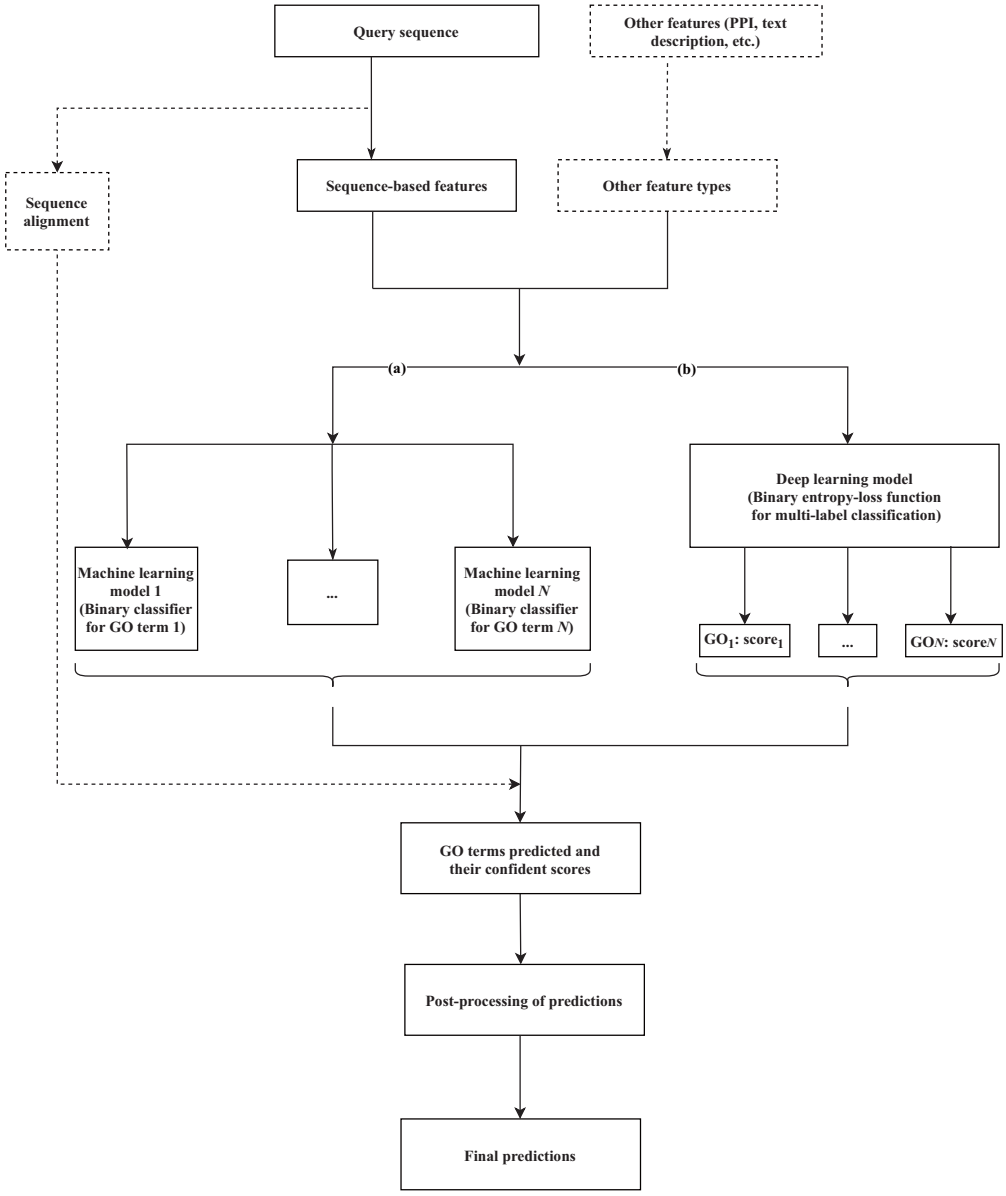


Figure 1 A common workflow of (A) machine learning-based and (B) deep learning-based solutions for predicting the GO terms of proteins. The components presented in the dashed box are optional, depending on each method. *N* is the number of GO terms that are selected for the training dataset to build the model.

Full-size DOI: 10.7717/peerj.12019/fig-1

Machine learning approaches are highly advantageous and considered the future direction for AFP. Some novel proteins either lack identifiable homologous sequences or their detectable homologues have not been assigned any GO labels. Consequently, assigning protein function from scratch using machine learning, *i.e.,* directly inferring the annotation from the amino acid sequence, without access to any additional references or databases,

is an ongoing task. The sophisticated algorithms are being developed, as presented in the next section.

GO TERM ANNOTATION OF PROTEINS USING DEEP LEARNING APPROACH

The excellent potential of deep learning—the latest discipline of machine learning—has been demonstrated in several application fields, including bioinformatics (*Min, Lee & Yoon, 2017*; *Li et al., 2019*). The main characteristic that distinguishes this approach from other methods is the learning process. Deep learning models are inspired by neural networks in the human brain. They automatically extract high-level features from raw data and provide predictions in an end-to-end manner. In contrast to that in conventional machine learning models, accurate classification is achieved *via* handcrafted features.

Currently, the amount of generated genomic data, and the numbers of sophisticated algorithms and computational resources are rapidly growing. These resources support deep learning to tackle the functional annotation problem. Below, we review proposed methods based on two criteria, *i.e.,* the model and the input data used. A summary table of the reviewed methods is presented in Table 2, and a common workflow for deep learning approach is illustrated as Figure 1B.

Model-based approach

Supervised learning

Supervised learning is an essential approach for *in silico* protein function prediction. Supervised models learn to annotate function guided by a training dataset, which is a group of characterized proteins with reliable, experimentally confirmed annotations. After the learning stage, a model that captures the relationship between feature and function is produced, and used to predict the GO terms for novel amino acid sequences.

Deep neural network (DNN) is a feed-forward architecture. It generally comprises an input layer, multiple hidden layers, and an output layer. Input data are processed in a unidirectional manner along the layers, from the first to the final stage. With respect to GO-based functional annotation, DNNs have been employed as multi-task DNNs (MTDNNs) (*Fa et al., 2018*) or a series of MTDNNs (*Rifaioglu et al., 2019*).

Convolutional neural network (CNN) was initially developed to handle two-dimensional images for handwritten digit recognition (*LeCun et al., 1990*). However, it has since become an effective architecture not only for multi-dimensional data, but also for one-dimensional input, such as sentences or genomic sequences. A CNN model starts with convolutional layers, whose units are feature maps. Each unit is obtained by computing convolution operations between local patches of a unit in the preceding layer and a filter (set of weights) (*LeCun, Bengio & Hinton, 2015*). Non-linearity pooling layers are stacked to convolutional layers, enabling the integration of different-scale features in this architecture. CNNs can be applied in AFP either alone (*Kulmanov, Khan & Hoehndorf, 2018*; *Cai, Wang & Deng, 2020*; *Du et al., 2020*) or in combination with other architectures (*Zhang et al., 2020*; *Spalević et al., 2020*).

Table 2 Deep learning-based methods for assigning GO terms to proteins.

Feature used	Program name	Description	DL model used
Sequence based	*Chicco, Sadowski & Baldi (2014)*	Deep AE-based method for GO term prediction.	AE
	Zou, Wang & Yu (2017)	Infers missing GO annotations for proteins of *Homo sapiens*, *S. cerevisiae*, *Mus musculus*, and *Drosophila* using deep restricted Boltzmann machines.	DRBM
	ProLanGO (*Cao et al. (2017)*	LSTM-based model using the idea of NMT for the functional annotation of amino acid sequences.	LSTM
	SECLEF (*Szalkai & Grolmusz, 2018b*)	Generates the GO function and UniProt protein family from sequence data. The core models were developed by *Szalkai & Grolmusz (2018a)*. The web-based interface is available at https://pitgroup.org/seclaf/.	CNN
	DEEPred (*Rifaioglu et al., 2019*)	GO-based protein function prediction with a set of connected MTDNN. The source code and data are available at https://github.com/cansyl/DEEPred.	DNN
	DeepSeq (*Nauman et al., 2019*)	Predicts *H. sapiens* protein function from sequence data only, using a deep learning architecture based on CNN. Source code and data available at https://github.com/recluze/deepseq.	CNN
	DeepGOPlus (*Kulmanov & Hoehndorf, 2020*)	Advanced method of DeepGO, relying on a structure of several CNN layers. Available at http://deepgoplus.bio2vec.net/.	CNN
Integrated data based/Structure based	*Tavanaei et al. (2016)*	Deep CNN model classifying functionality, based on the tertiary structure of human proteins.	CNN
	DeepGO (*Kulmanov, Khan & Hoehndorf, 2018*)	Deep learning framework, employing one convolutional layer. Protein GO annotations are inferred based on a hierarchically structured classifier. Online tool available at https://deepgo.cbrc.kaust.edu.sa/deepgo/.	CNN
	deepNF (*Gligorijević, Barot & Bonneau, 2018*)	Deep network fusion, captures high-level features of several network data for AFP. Source code and data available at https://github.com/VGligorijevic/deepNF.	AE
	(*Fa et al., 2018*)	An implementation of two stacked multi-layer structures to predict GO terms from the sequence and structure properties. Source code and data available at http://bioinf.cs.ucl.ac.uk/downloads/mtdnn.	DNN
	DeepFunc (*Zhang et al., 2019*)	Deduces Go annotations by concatenating InterPro-based and protein interaction network features.	FCDN
	DeepGOA (*Zhang et al., 2020*)	Integrates comprehensive sequence and PPI information for AFP.	Bi-LSTM, CNN
	SDN2GO (*Cai, Wang & Deng, 2020*)	Integrates architectures with three sub-models and a weighting classifier to achieve GO term predictions. Source code and data available at https://github.com/Charrick/SDN2GO.	CNN
	DeepAdd (*Du et al., 2020*)	CNN-based framework, predicting protein function from sequence and additional information (PPI or SSP).	CNN
	FFPred-GAN (*Wan & Jones, 2020*)	Novel FFPred that uses training sample augmented by a deep learning architecture.	GAN
	GONET (*Li et al., 2020*)	Deep model for protein function prediction using representation learning to embed protein sequences and networks.	CNN, RNN

Recurrent neural network (RNN) is a deep learning architecture developed especially for sequential data. RNNs have a DNN backbone, and the units of the hidden layer are connected. Therefore, a hidden unit receives information from the input layer at the current time step, and also from the hidden units at the preceding stage (*LeCun, Bengio & Hinton, 2015*). Long short-term memory (LSTM) (*Hochreiter & Schmidhuber, 1997*) is a variant of the RNN model, developed to capture long-term dependencies between input sequences. Bi-directional LSTM (Bi-LSTM) is an LSTM model that processes data in two directions, forwards and backwards. These neural networks are employed to process protein sequences as a natural language (*Cao et al., 2017*) or are combined with the CNN model (*Zhang et al., 2020*) to provide GO annotations.

Another type of neural network, fully connected deep network (FCDN), is a series of fully connected layers. This model was utilized for converting input vectors from InterPro and predicting GO terms (*Zhang et al., 2019*).

Unsupervised learning

Unlike supervised learning, unsupervised learning independently mines hidden patterns in input distribution. It is crucial for exploring unlabeled data and provides important information for supervised architectures. Typically, unsupervised learning models are employed for clustering, reducing dimensions, and transforming data.

Autoencoder (AE) (*Baldi, 2012*) is an unsupervised learning model developed for assigning GO terms to amino acid sequences. AE is a neural network for data transformation. After encoding and decoding, the target output size is the same as that of the input. Using a trained AE model, GO predictions for proteins can be deduced directly from the generated matrix (*Chicco, Sadowski & Baldi, 2014*). Alternatively, low-dimensional feature representations in the hidden layer are extracted and fed into an SVM classifier for final classification (*Gligorijević, Barot & Bonneau, 2018*) or a CNN model (*Peng et al., 2020*) for final classification.

Restricted Boltzmann machine (RBM) (*Salakhutdinov & Hinton, 2009*) has one hidden layer for representing latent features and an input layer encoding the observed data. For GO annotation, units of the input layer are the available GO terms (*Zou, Wang & Yu, 2017*). *Zou, Wang & Yu (2017)* introduced a deep RBM (DRBM), where multi-layer RBMs are trained and unfolded, resulting in protein function prediction.

Finally, generative adversarial network (GAN) (*Goodfellow et al., 2014*) comprises two networks, *i.e.,* a generative model that produces synthetic realistic data and a discriminative model that evaluates whether the data are real or not. In AFP, the GAN model improves the prediction by creating synthetic features for supervised classifiers (*Wan & Jones, 2020*), or by exploiting ontology correlations (*Seyyedsalehi et al., 2021*).

Data-based approach

In terms of the input feature, deep learning-based methods primarily use two approaches to assign GO terms to unknown gene products. One is a sequence-only approach, which is useful for predicting the functions of novel proteins in the absence of homologous information or other references. The second one is structure-based or otherwise exploits big data from several available resources.

Sequence-based models

One of the earliest deep learning-based methods for GO annotation predictions is the one proposed by *Chicco, Sadowski & Baldi (2014)*. The authors compared two annotation solutions, *i.e.,* tSVD, presented in *Pinoli, Chicco & Masseroli (2013)*, and the AE neural network. The output form of the two methods was similar; however, the latter performed better on six different datasets. Later, instead of using the AE architecture, (*Zou, Wang & Yu, 2017*) proposed DRBM for annotating the gene products of four model species, *Homo sapiens, S.cerevisiae, Musmusculus*, and *Drosophila*.

ProLanGo (*Cao et al., 2017*) is the first tool that applies Neural Machine Translation (NMT) developed by Google in AFP. The authors converted amino acid sequences and GO terms into "ProLan" and "GOLan" languages, respectively. The GO annotation of a new protein was then generated after being translated using a model with three RNN layers. In another study, artificial neural networks proposed by *Szalkai & Grolmusz (2018a)* utilized six CNN layers of increasing depth. Based on that work, SECLEF (*Szalkai & Grolmusz, 2018b*) has been designed to train and test the GO function and UniProt protein family using biological sequences as input. The user can implement their self-configured models in a downloadable program or run two pre-trained models (*Szalkai & Grolmusz, 2018a*) on a web server. DeepSeq (*Nauman et al., 2019*) is another method for assigning GO terms to amino acid sequences, based on convolutional layers. However, the authors only predicted the five most frequent MF ontologies for *H. sapiens* proteins.

Using the DNN architecture, DEEPred (*Rifaioglu et al., 2019*) implements a stack of multi-task feed-forward DNNs. Each individual network is built for a specific GO term level on DAG, which allows the hierarchical post-processing of predictions. Protein sequences are experimentally represented by three types of descriptors(subsequence profile map, pseudo-amino acid composition, and conjoint triad feature), with subsequence profile map performing best in the analysis.

DeepGOPlus (*Kulmanov & Hoehndorf, 2020*) has been developed to overcome the existing limitations of DeepGO (*Kulmanov, Khan & Hoehndorf, 2018*), such as sequence length, unavailable PPI features, and the number of GO labels. Functional annotation is predicted by a multi-layer CNN structure in combination with sequence similarity. Lately, TALE (*Cao & Shen, 2021*) has been developed to generate GO predictions by integrating sequence patterns based on transformer encoder and the joint similarity of sequence-term. Similar to DeepGOPlus, TALE is also combined with sequence similarity as TALE+ model to enhance its performance. Meanwhile, PFP-WGAN (*Seyyedsalehi et al., 2021*) is one of two latest ideas that use GAN to infer the functionalities of proteins. While the generator network processes raw sequences, the discriminator takes two inputs; one consists of annotated proteins from the SwissProt database and the other of raw sequences with respective annotations synthesized from the generator.

Integrated data-based/Structure based-models

The functional assignment of amino acid sequences from 3D structures was proposed by (*Tavanaei et al., 2016*). The main analysis involved a CNN model that was trained and

tested on five datasets of human proteins, each of which had been assigned two GO terms. However, the prediction has not been propagated on the DAG tree for inherited GO terms.

DeepGO (*Kulmanov, Khan & Hoehndorf, 2018*) is a functional annotation server based on CNN architecture that had been devised to learn features from amino acid sequences and the PPI network. GO labels are assigned to input proteins through a hierarchical classification structured as a DAG tree. DeepAdd (*Du et al., 2020*) was inspired by DeepGO server and provides a solution for AFP, utilizing a CNN framework to learn vector representations from sequences and additional information. However, sequences are processed into k-mer embedding by Word2Vec, instead of the tri-gram embedding suggested in DeepGO. Further, in DeepAdd, the protein sequence profile (SSP) is added if the PPI network information is not available. Based on a similar concept of combining primary protein structure and PPI, GONET (*Li et al., 2020*), a novel model, was built by employing CNN, RNN, and Attention layer for human and mouse sequences.

Working on a dataset identical to the one collated by FFPred3, *Fa et al. (2018)* tested an MTDNN that treated protein function prediction as a multi-label classification problem. Here, the solution consisted of layers shared by all tasks (GO labels), which are stacked in parallel with task-specific layers. DeepFunc (*Zhang et al., 2019*) is a novel predictor that surpasses DeepGO, FFPred3, and BLAST. First, protein domain, family, and motif information is queried from InterPro and encoded before passing through fully connected layers. Next, topological features of PPI are obtained by the Deepwalk algorithm. Finally, the method concatenates two types of features (sequence- and network-based input) to fit an FCDN. The architecture of DeepGOA (*Zhang et al., 2020*) is more sophisticated than that of DeepFunc. In addition to information generated by InterPro and PPI, global and local semantic features of amino acid sequences are extracted by Bi-LSTM and a convolutional layer, respectively. Using the same types of features, SDN2GO (*Cai, Wang & Deng, 2020*) utilizes three sub-models for each information source, with all outputs integrated in the final weighted model.

In another study, deepNF (*Gligorijević, Barot & Bonneau, 2018*) was constructed by multimodal deep AE to capture hidden information in proteins from different types of interaction networks. In that suggestion, feature representations are extracted to train the final SVM classifier to produce GO terms for human proteins. DeepMNE-CNN (*Peng et al., 2020*) has a superior performance than deepNF in the human data by utilizing CNN layers instead of SVM for the classification model. On the other hand, FFPred-GAN (*Wan & Jones, 2020*) utilizes GAN for feature enrichment to feed conventional machine learning models, especially SVM. Real features are biophysical information extracted from raw amino acid sequences by FFPred, and the generator takes latent variables to augment the synthetic samples.

COMPARISON

In addition to proposing a new approach, accuracy assessment is important for demonstrating the improvements of a novel methodology. This necessitates the establishment of a common framework for the evaluation and comparison of the proposed

Table 3 Statistics of final benchmark CAFA3.

Annotation type	Number of sequences	Number of terms (GO leaf only)		
		MF	CC	BP
NK (no-knowledge)	1177	377	236	1087
LK (limited-knowledge)	2151	368	203	878
All (NK + LK)	3328	631	327	1673

solutions. Accordingly, the critical assessment of functional annotation (CAFA) (*Radivojac et al., 2013*; *Jiang et al., 2016*; *Zhou et al., 2019*) is a community-based experiment that provides a large-scale evaluation of computational protein function prediction methods in a time-delayed manner. Four main competitions (CAFA1–CAFA4) have been held every 3 years since 2010. CAFA-π is an extension of CAFA3, and allows the competitors to improve their computational solutions focusing on the prediction of genes associated with specific GO terms. Generally, each challenge comprises three phases, *i.e.,* predictions for a set of target proteins, experimental verification for a subset of target proteins to obtain a benchmark dataset, and performance analysis of all the competitors based on predefined metrics and the benchmark data. Official reports have just been released for the first three challenges and CAFA-π (*Radivojac et al., 2013*; *Jiang et al., 2016*; *Zhou et al., 2019*).

In the remainder of this section, we present a practical comparison of emerging analysis methods and discuss the retrieved results.

Set up

Data

We used the finalized benchmark set provided in the CAFA3 report (*Zhou et al., 2019*). The statistics are presented in Table 3. There are two sequence types in the benchmark dataset, NK (no-knowledge) and LK (limited-knowledge), which have been coined in CAFA2 (*Jiang et al., 2016*). Based on the timeline of the challenge, t_0 is the deadline for the prediction submission of the target sequences. Then, experimental annotations for a subset of target sequences are accumulated until t_1, to complete a benchmark dataset for the performance evaluation. The NK proteins are those that had not had any experimental annotations in three GO domains at t_0, but had at least one GO term with experimental evidence code between t_0 and t_1. The LK sequences are those that had been experimentally annotated in one or two GO domains at t_0, and gained at least one experimental annotation in the other GO domains between t_0 and t_1.

Methods

The servers of several tools have been updated recently, but we need to control their training data or databases used, in which do not include sequences in the CAFA3 benchmark. Therefore, we chose INGA, which was one of the top models in CAFA3, as a representative conventional tool. DeepGOPlus is a recently developed representative annotation system that is based on deep learning. In addition to these two tools, other studies that performed their analysis on the CAFA3 dataset are also summarized in this part.

For DeepGOPlus, we re-trained the model with the CAFA3 training sequences and generated predictions for the CAFA3 benchmark set based on their source code provided. For INGA, we obtained predictions for the CAFA3 benchmark based on the predictions of CAFA3 targets, which were provided on their website.

Model performances were evaluated based on protein-centric metrics (F_{max}, precision, and recall) provided by the CAFA3 assessment tool. The area under the precision–recall curve (AUPR), which is a measurement of highly imbalanced data, was also considered. Results regarding the performance of other methods, including Naive, BLAST, FFPred3 and DEEPred, were derived from DEEPred paper (*Rifaioglu et al., 2019*).

In terms of evaluation, two modes were designed. In the *partial* mode, the scores of each method were computed on the sequences that have at least one GO term predicted at any threshold. Conversely, scores in the *full* mode were computed on all protein sequences in the benchmark set.

Results

The results of the analysis are shown in terms of GO domains and types of sequences (NK *vs.* All) in *full* mode (Fig. 2) and *partial* mode (Fig. 3). The scores of Naive, BLAST, FFPred3 and DEEPred are only available for the *full* mode. Models using the traditional approach and deep learning approach are depicted in blue and orange tones, respectively.

As is shown in Fig. 2, deep learning-based models (DEEPred and DeepGOPlus) showed a competitive performance compared to conventional models in terms of MF and BP prediction; in particular, DEEPred acquired the highest precision in the three GO categories. Such high precision obtained by DEEPred, which was explained in their paper (*Rifaioglu et al., 2019*), probably resulted from the imbalanced dataset and post-processing step. For example, the negative protein set for each GO term is four times as large as the positive one, this leads to low false positive and high false negative predictions. Additionally, DEEPred filtered some false positives at the post-processing step, thus increasing their precision. For the CC class, two baselines (Naive and BLAST) were at the top for F_{max} and recall, followed by FFPred3. The overall outcome is presented similarly in Fig. 3, in which DeepGOPlus performed better than INGA in most of categories.

The aforementioned results and the CAFA3 report indicate that conventional models, such as INGA and GOLabeler (referred to as "ZhuLab" in CAFA3) perform as well as other high-performing models. ZhuLab was the best performer in terms of F_{max} and NK-*full* evaluation, with the scores of 0.40, 0.61, and 0.62 for BP, CC, and MF, respectively. However, these conventional models integrate several types of information, which is not always available for all sequences. Meanwhile, the deep learning approach had demonstrable advantages in this area and yielded highly competitive predictions. The research direction is currently shifting toward this new approach, with even more efficient solutions being developed.

CONCLUSION

The computational GO annotation of proteins has been an actively pursued and challenging task in bioinformatics since around the 2000s; this is a response to the need to bridge the

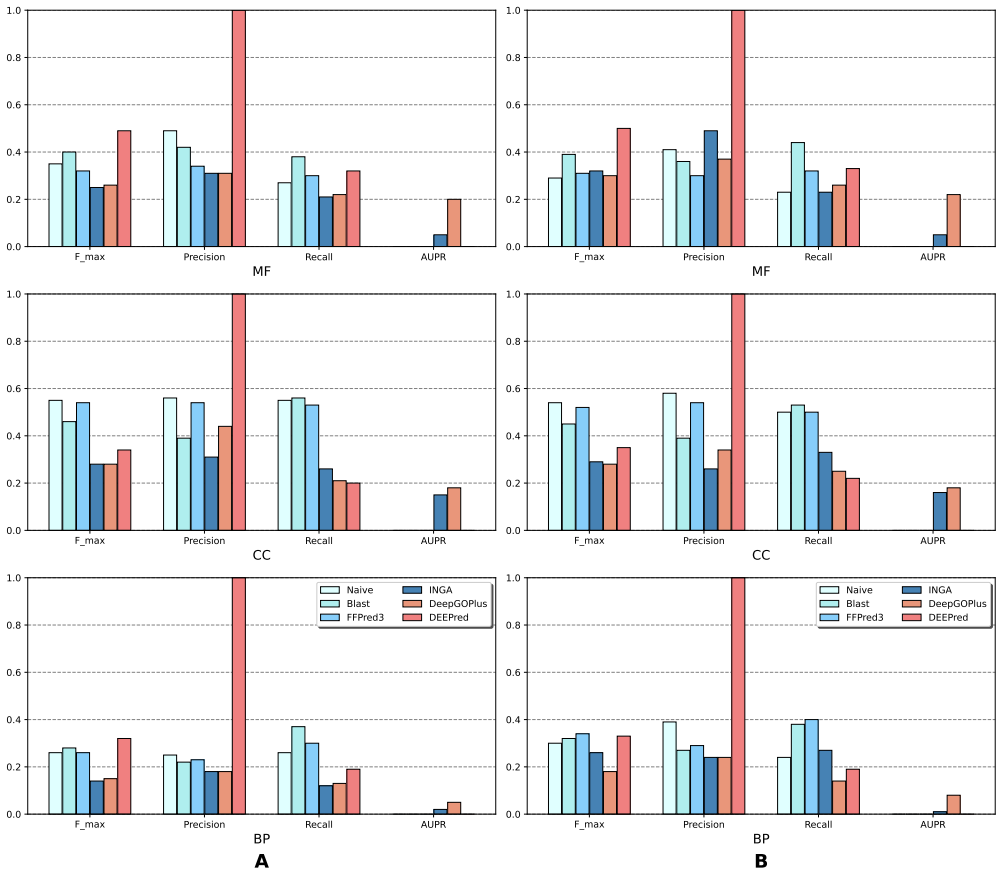


Figure 2 Performance of the selected methods with the CAFA3 benchmark dataset in *full* mode: (A) NK sequences and (B) All sequences (NK + LK). Scores in the *full* mode were computed on all protein sequences in the benchmark set. Scores of Naive, BLAST, FFPred3, DEEPred have been reported directly from *Rifaioglu et al. (2019)*.

Full-size DOI: 10.7717/peerj.12019/fig-2

gap between the known and unknown, newly discovered amino acid sequences. On the one hand, understanding protein function is essential for deciphering biological evolution and for countless applications, such as drug design and disease treatment. The GO database has facilitated a comprehensive vocabulary for functional annotation because it presents structured GO functions in three domains (MF, BP, and CC), thereby effectively supporting *in silico* protein function assignment. Herein, we presented the current state of the field and compared GO-based AFP solutions, classed as conventional and deep learning approaches. As it is difficult to introduce all the available methodologies and comprehensively compare them, we tested the prominent predictors in this study. This was complemented by the analysis of related publications, with the assessment outcomes computed by CAFA, a worldwide venue for comparing computational protein function predictors.

We observed that one of the primary difficulties in the field is the utilization of input features to achieve effective performance. While heterogeneous input data are useful for GO prediction, only sequence information is available for the majority of unannotated

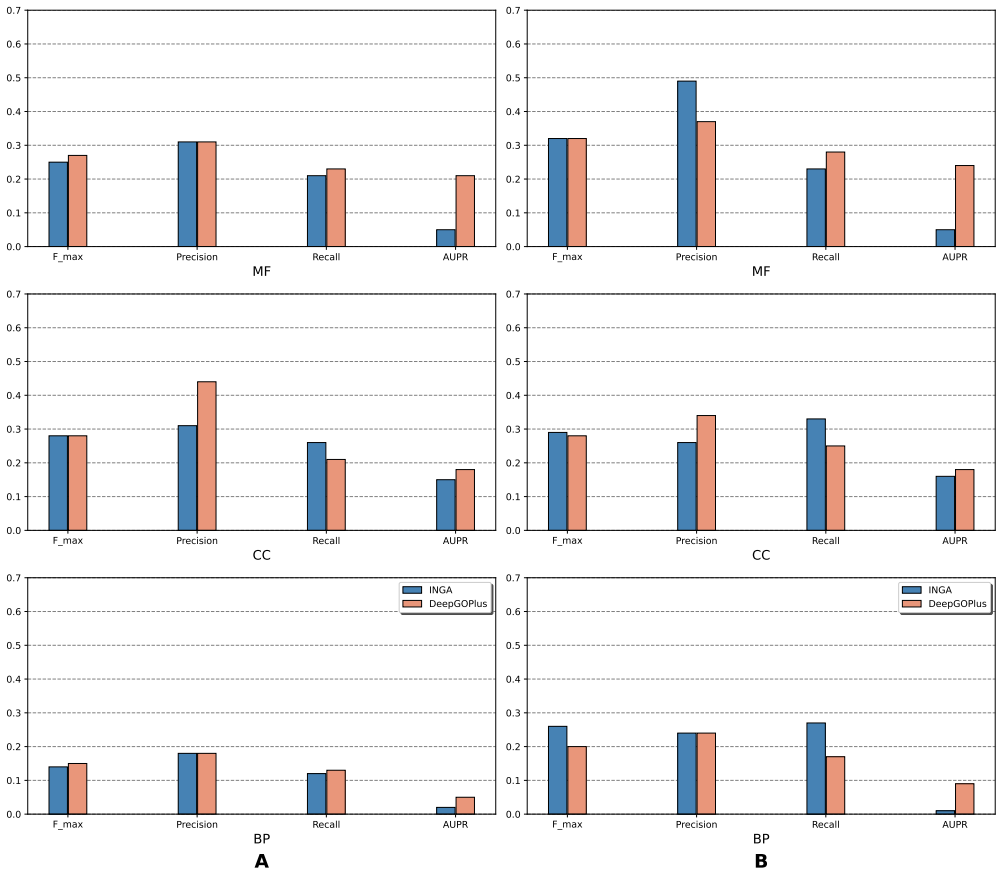


Figure 3 Performance of the selected methods with the CAFA3 benchmark dataset in *partial* mode: (A) NK sequences and (B) all sequences (NK + LK). In the *partial* mode, scores of each method were computed on the sequences that have at least one GO term predicted at any threshold.
Full-size DOI: 10.7717/peerj.12019/fig-3

proteins. With respect to the specific approaches, machine learning models may be limited by the heterogeneity of genomic data when mining different sources of information, while parameter and hyper-parameter tuning is the challenging step of the deep learning approach. In terms of output, the GO database has been updating because GO annotations are still imbalanced and not complete for all species. Its sophisticated structure is ideal for describing the functional roles of proteins, but also makes the prediction task a complex multi-label problem.

Nonetheless, despite the existing challenges, the quest for successful annotation will continue in many directions, driven by the community effort. Together with the constant expansion of -omics databases, data-driven methods have shown excellent applications in various areas, with a promising trend in functional annotation. Depending on the available sources, they can be based on integrated data or sequence information only, considering a hybrid approach. Some approaches have been suggested and developed to cope with imbalanced data; these include working on the subgroups in which the classes are more balanced, data augmentation using GAN, and considering evaluation metrics specified

for imbalanced data, such as AUPR. The multi-label problem can be addressed *via* the advancement of computational resources and well-defined solutions, for example, stacking many individual solutions together. Finally, successful solutions for GO term prediction can be expanded to include other functional resources (EC, pathways, etc.), to comprehend the biological role and life science potential of proteins.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2019R1A2C1084308). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning: NRF-2019R1A2C1084308.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Thi Thuy Duong Vu conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Jaehee Jung conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Supplemental documents are available at GitHub: https://github.com/duongvtt96/Comparison-GO-annotation-systems

Dataset used in this comparison (final benchmark CAFA3) is available at Figshare:

Zhou, Naihui (2019): Supplementary_data. figshare. Dataset. https://doi.org/10.6084/m9.figshare.8135393.v3.

REFERENCES

Anonymous. 2018. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **46(D1)**:D8–D13 DOI 10.1093/nar/gkx1095.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402 DOI 10.1093/nar/25.17.3389.

Amiri-Dashatan N, Koushki M, Abbaszadeh H-A, Rostami-Nejad M, Rezaei-Tavirani M. 2018. Proteomics applications in health: biomarker and drug discovery and food industry. *Iranian journal of pharmaceutical research: IJPR* **17(4)**:1523–1536.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Baldi P. 2012. Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 37–49.

Bonetta R, Valentino G. 2020. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics* **88(3)**:397–413 DOI 10.1002/prot.25832.

Cai Y, Wang J, Deng L. 2020. Sdn2go: an integrated deep learning model for protein function prediction. *Frontiers in Bioengineering and Biotechnology* **8**:391 DOI 10.3389/fbioe.2020.00391.

Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. 2017. Prolango: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* **22(10)**:1732 DOI 10.3390/molecules22101732.

Cao Y, Shen Y. 2021. TALE: transformer-based protein function Annotation with joint sequence–Label Embedding. *Bioinformatics* btab198 DOI 10.1093/bioinformatics/btab198.

Chicco D, Sadowski P, Baldi P. 2014. Deep autoencoder neural networks for gene ontology annotation predictions. In: *Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics*. 533–540.

Consortium GO. 2015. Gene ontology consortium: going forward. *Nucleic Acids Research* **43(D1)**:D1049–D1056 DOI 10.1093/nar/gku1179.

Cozzetto D, Jones DT. 2017. Computational methods for annotation transfers from sequence. In: *The gene ontology handbook*. New York: Humana Press, 55–67.

Cozzetto D, Minneci F, Currant H, Jones DT. 2016. Ffpred 3: feature-based function prediction for all gene ontology domains. *Scientific Reports* **6(1)**:1–11 DOI 10.1038/s41598-016-0001-8.

Deng M, Chen T, Sun F. 2004a. An integrated probabilistic model for functional prediction of proteins. *Journal of Computational Biology* **11(2–3)**:463–475 DOI 10.1089/1066527041410346.

Deng M, Tu Z, Sun F, Chen T. 2004b. Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics* **20(6)**:895–902 DOI 10.1093/bioinformatics/btg500.

Du Z, He Y, Li J, Uversky VN. 2020. Deepadd: protein function prediction from k-mer embedding and additional features. *Computational Biology and Chemistry* **89**:107379 DOI 10.1016/j.compbiolchem.2020.107379.

Fa R, Cozzetto D, Wan C, Jones DT. 2018. Predicting human protein function with multi-task deep neural networks. *PLOS ONE* **13(6)**:e0198216 DOI 10.1371/journal.pone.0198216.

Freund Y, Schapire RE. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55(1)**:119–139 DOI 10.1006/jcss.1997.1504.

Friedberg I. 2006. Automated protein function prediction—the genomic challenge. *Briefings in Bioinformatics* **7(3)**:225–242 DOI 10.1093/bib/bbl004.

Gligorijević V, Barot M, Bonneau R. 2018. deepnf: deep network fusion for protein function prediction. *Bioinformatics* **34(22)**:3873–3881 DOI 10.1093/bioinformatics/bty440.

Gong Q, Ning W, Tian W. 2016. Gofdr: a sequence alignment based method for predicting protein functions. *Methods* **93**:3–14 DOI 10.1016/j.ymeth.2015.08.009.

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial networks. ArXiv preprint. arXiv:1406.2661.

Hawkins T, Chitale M, Luban S, Kihara D. 2009. Pfp: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics* **74(3)**:566–582 DOI 10.1002/prot.22172.

Hennig S, Groth D, Lehrach H. 2003. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Research* **31(13)**:3712–3715 DOI 10.1093/nar/gkg582.

Hewelt A, Ben Kahla A, Hennig S, Nagel A, Himmelbauer H, Zehetner G, Haas S, Vingron M, Yaspo M, Lehrach H. 2002. The genomematrix information retrieval system. In: *Poster Abstracts of HGM2002 Human Genome Meeting (HGM2002)*. Shanghai, China: HUGO.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation* **9(8)**:1735–1780 DOI 10.1162/neco.1997.9.8.1735.

Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo DCE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SME, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Törönen P, Koskinen P, Holm L, Chen C-T, Hsu W-L, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Dukka BKC, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent L-C, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeño-Cortés AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SCE, Del Pozo A, Fernández JM, Maietta P, Valencia A,

Tress ML, Benso A, Carlo SD, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, Van Dijk ADJ, Gemovic B, Glisic S, Perovic V, Veljkovic Veljko, Veljkovic N, Almeida-e-Silva DC, Vencio RZN, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z , et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* **17(1)**:1–19 DOI 10.1186/s13059-015-0866-z.

Jung J, Thon MR. 2006. Automatic annotation of protein functional class from sparse and imbalanced data sets. In: *VLDB workshop on data mining and bioinformatics*. Springer, 65–77.

Jung J, Yi G, Sukno SA, Thon MR. 2010. Pogo: prediction of gene ontology terms for fungal proteins. *BMC Bioinformatics* **11(1)**:215 DOI 10.1186/1471-2105-11-215.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The kegg resource for deciphering the genome. *Nucleic Acids Research* **32(suppl_1)**:D277–D280 DOI 10.1093/nar/gkh063.

Khan S, Situ G, Decker K, Schmidt CJ. 2003. Gofigure: automated gene ontology annotation. *Bioinformatics* **19(18)**:2484–2485 DOI 10.1093/bioinformatics/btg338.

Kourmpetis YA, Van Dijk AD, Bink MC, Van Ham RC, ter Braak CJ. 2010. Bayesian markov random field analysis for protein function prediction based on network data. *PLOS ONE* **5(2)**:e9293 DOI 10.1371/journal.pone.0009293.

Kulmanov M, Hoehndorf R. 2020. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics* **36(2)**:422–429.

Kulmanov M, Khan MA, Hoehndorf R. 2018. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34(4)**:660–668 DOI 10.1093/bioinformatics/btx624.

Lan L, Djuric N, Guo Y, Vucetic S. 2013. Ms-k nn: protein function prediction by integrating multiple data sources. *BMC Bioinformatics* **14**:S8.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521(7553)**:436–444 DOI 10.1038/nature14539.

LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD. 1990. Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*. Denver, Colorado, USA: Morgan Kaufmann, 396–404.

Letovsky S, Kasif S. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19(suppl_1)**:i197–i204 DOI 10.1093/bioinformatics/btg1026.

Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. 2019. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods* **166**:4–21 DOI 10.1016/j.ymeth.2019.04.008.

Li J, Wang L, Zhang X, Liu B, Wang Y. 2020. Gonet: a deep network to annotate proteins via recurrent convolution networks. In: *2020 IEEE international conference on bioinformatics and biomedicine (BIBM), volume 2*. Piscataway: IEEE, 29–34.

Lobley AE, Nugent T, Orengo CA, Jones DT. 2008. Ffpred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Research* **36(suppl_2)**:W297–W302 DOI 10.1093/nar/gkn193.

Lobley A, Swindells MB, Orengo CA, Jones DT. 2007. Inferring function using patterns of native disorder in proteins. *PLOS Computational Biology* **3(8)**:e162 DOI 10.1371/journal.pcbi.0030162.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402(6757)**:83–86 DOI 10.1038/47048.

Martin DM, Berriman M, Barton GJ. 2004. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* **5(1)**:178 DOI 10.1186/1471-2105-5-178.

Min S, Lee B, Yoon S. 2017. Deep learning in bioinformatics. *Briefings in Bioinformatics* **18(5)**:851–869.

Nariai N, Kolaczyk ED, Kasif S. 2007. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLOS ONE* **2(3)**:e337 DOI 10.1371/journal.pone.0000337.

Nauman M, Rehman HU, Politano G, Benso A. 2019. Beyond homology transfer: deep learning for automated annotation of proteins. *Journal of Grid Computing* **17(2)**:225–237 DOI 10.1007/s10723-018-9450-6.

Pandey G, Kumar V, Steinbach M. 2006. *Computational approaches for protein function prediction: a survey*. Twin Cities: Department of Computer Science and Engineering, University of Minnesota, 1.

Peng J, Xue H, Wei Z, Tuncali I, Hao J, Shang X. 2020. Integrating multi-network topology for gene function prediction using deep neural networks. *Briefings in Bioinformatics* **22(2)**:2096–2105.

Pinoli P, Chicco D, Masseroli M. 2013. Improved biomolecular annotation prediction through weighting scheme methods. In: *International meeting on computational intelligence methods for bioinformatics and biostatistics, Tenth edition, CIBB 2013*. Universitá di Salerno, 1–12.

Pinoli P, Chicco D, Masseroli M. 2015. Computational algorithms to predict gene ontology annotations. *BMC Bioinformatics* **16(6)**:1–15.

Piovesan D, Giollo M, Leonardi E, Ferrari C, Tosatto SC. 2015. Inga: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Research* **43(W1)**:W134–W140 DOI 10.1093/nar/gkv523.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P,

Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJE, Škunca N, Supek F, Bošnjak M, Panov P, Džeroski S, Šmuc T, Kourmpetis YAI, Van Dijk ADJ, ter Braak CJF, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I. 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods* **10(3)**:221–227 DOI 10.1038/nmeth.2340.

Rentzsch R, Orengo CA. 2009. Protein function prediction–the power of multiplicity. *Trends in Biotechnology* **27(4)**:210–219 DOI 10.1016/j.tibtech.2009.01.002.

Rifaioglu AS, Doğan T, Martin MJ, Cetin-Atalay R, Atalay V. 2019. Deepred: automated protein function prediction with multi-task feed-forward deep neural networks. *Scientific Reports* **9(1)**:1–16.

Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. 2003. Automatic prediction of protein function. *Cellular and Molecular Life Sciences* **60(12)**:2637–2650 DOI 10.1007/s00018-003-3114-8.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW. 2004. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research* **32(18)**:5539–5545 DOI 10.1093/nar/gkh894.

Salakhutdinov R, Hinton G. 2009. Deep boltzmann machines. In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Clearwater, Florida, USA: PMLR, 448–455.

Sasson O, Kaplan N, Linial M. 2006. Functional annotation prediction: all for one and one for all. *Protein Science* **15(6)**:1557–1562 DOI 10.1110/ps.062185706.

Seyyedsalehi SF, Soleymani M, Rabiee HR, Mofrad MR. 2021. Pfp-wgan: protein function prediction by discovering gene ontology term correlations with generative adversarial networks. *PLOS ONE* **16(2)**:e0244430 DOI 10.1371/journal.pone.0244430.

Sharan R, Ulitsky I, Shamir R. 2007. Network-based prediction of protein function. *Molecular Systems Biology* **3(1)**:88 DOI 10.1038/msb4100129.

Shehu A, Barbará D, Molloy K. 2016. A survey of computational methods for protein function prediction. Cham, Switzerland: Springer International Publishing, 225–298.

Sleator RD, Walsh P. 2010. An overview of in silico protein function prediction. *Archives of Microbiology* **192(3)**:151–155 DOI 10.1007/s00203-010-0549-9.

Spalević S, Veličković P, Kovačević J, Nikolić M. 2020. Hierachial protein function prediction with tails-gnns. ArXiv preprint. arXiv:2007.12804.

Stein L. 2001. Genome annotation: from sequence to biology. *Nature Reviews Genetics* **2(7)**:493–503.

Szalkai B, Grolmusz V. 2018a. Near perfect protein multi-label classification with deep neural networks. *Methods* **132**:50–56 DOI 10.1016/j.ymeth.2017.06.034.

Szalkai B, Grolmusz V. 2018b. Seclaf: a webserver and deep neural network design tool for hierarchical biological sequence classification. *Bioinformatics* **34(14)**:2487–2489 DOI 10.1093/bioinformatics/bty116.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, Von Mering C. 2016. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45(D1)**:D362–D368.

Tavanaei A, Maida AS, Kaniymattam A, Loganantharaj R. 2016. Towards recognition of protein function based on its structure using deep convolutional networks. In: *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. Piscataway: IEEE, 145–149.

Törönen P, Medlar A, Holm L. 2018. Pannzer2: a rapid functional annotation web server. *Nucleic Acids Research* **46(W1)**:W84–W88 DOI 10.1093/nar/gky350.

Vinayagam A, del Val C, Schubert F, Eils R, Glatting K-H, Suhai S, König R. 2006. Gopet: a tool for automated predictions of gene ontology terms. *BMC Bioinformatics* **7(1)**:161 DOI 10.1186/1471-2105-7-161.

Wan C, Jones DT. 2020. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nature Machine Intelligence* **2(9)**:540–550 DOI 10.1038/s42256-020-0222-1.

Watson JD, Laskowski RA, Thornton JM. 2005. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology* **15(3)**:275–284 DOI 10.1016/j.sbi.2005.04.003.

Webb EC, editor. 1992. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* San Diego: Academic Press.

You R, Huang X, Zhu S. 2018a. Deeptext2go: improving large-scale protein function prediction with deep semantic text representation. *Methods* **145**:82–90 DOI 10.1016/j.ymeth.2018.05.026.

You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S. 2019. Netgo: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research* **47(W1)**:W379–W387 DOI 10.1093/nar/gkz388.

You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. 2018b. Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* **34(14)**:2465–2473 DOI 10.1093/bioinformatics/bty130.

Zdobnov EM, Apweiler R. 2001. Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics* **17(9)**:847–848 DOI 10.1093/bioinformatics/17.9.847.

Zehetner G. 2003. Ontoblast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research* **31(13)**:3799–3803 DOI 10.1093/nar/gkg555.

Zhang F, Song H, Zeng M, Li Y, Kurgan L, Li M. 2019. Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics* **19(12)**:1900019 DOI 10.1002/pmic.201900019.

Zhang F, Song H, Zeng M, Wu F-X, Li Y, Pan Y, Li M. 2020. A deep learning framework for gene ontology annotation with sequence-and network-based information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* Epub ahead of print Jan 23 2020 DOI 10.1109/TCBB.2020.2968882.

Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MN , et al. 2019. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology* **20(1)**:1–23 DOI 10.1186/s13059-018-1612-0.

Zhao Y, Wang J, Chen J, Zhang X, Guo M, Yu G. 2020. A literature review of gene function prediction by modeling gene ontology. *Frontiers in Genetics* **11**:400 DOI 10.3389/fgene.2020.00400.

Zou X, Wang G, Yu G. 2017. Protein function prediction using deep restricted boltzmann machines. *BioMed Research International* **2017**:1729301.

Human Genetics (2024) 143:625–634
https://doi.org/10.1007/s00439-024-02670-5

REVIEW

Variant effect predictors: a systematic review and practical guide

Cristian Riccio[1,2] · Max L. Jansen[1,2] · Linlin Guo[3,4] · Andreas Ziegler[1,2,3,4,5]

Received: 13 December 2023 / Accepted: 11 March 2024 / Published online: 4 April 2024
© The Author(s) 2024

Abstract

Large-scale association analyses using whole-genome sequence data have become feasible, but understanding the functional impacts of these associations remains challenging. Although many tools are available to predict the functional impacts of genetic variants, it is unclear which tool should be used in practice. This work provides a practical guide to assist in selecting appropriate tools for variant annotation. We conducted a MEDLINE search up to November 10, 2023, and included tools that are applicable to a broad range of phenotypes, can be used locally, and have been recently updated. Tools were categorized based on the types of variants they accept and the functional impacts they predict. Sequence Ontology terms were used for standardization. We identified 118 databases and software packages, encompassing 36 variant types and 161 functional impacts. Combining only three tools, namely SnpEff, FAVOR, and SparkINFERNO, allows predicting 99 (61%) distinct functional impacts. Thirty-seven tools predict 89 functional impacts that are not supported by any other tool, while 75 tools predict pathogenicity and can be used within the ACMG/AMP guidelines in a clinical context. We launched a website allowing researchers to select tools based on desired variants and impacts. In summary, more than 100 tools are already available to predict approximately 160 functional impacts. About 60% of the functional impacts can be predicted by the combination of three tools. Unexpectedly, recent tools do not predict more impacts than older ones. Future research should allow predicting the functionality of so far unsupported variant types, such as gene fusions.
URL: https://cardio-care.shinyapps.io/VEP_Finder/.
Registration: OSF Registries on November 10, 2023, https://osf.io/s2gct.

Introduction

Whole-genome sequencing (WGS) has become precise and affordable on a large scale, and several cohorts now involve hundreds of thousands of subjects (Halldorsson et al. 2022; Taub et al. 2022; The All of Us Research Program Investigators 2019). Statistical analyses have associated diseases with common and rare variants (Povysil et al. 2019), and the GWAS Catalog currently contains more than half a million associations (Sollis et al. 2023). However, the causal mechanism behind most genetic associations is unclear and can take a long time to understand. For example, even for the best-replicated locus in cardiovascular disease, it took four years to unravel its function (Harismendy et al. 2011; Wellcome Trust Case Control Consortium 2007). To accelerate the understanding of biological function, a series of computational tools have been proposed in recent years.

In this work, we consider Variant Effect Predictors (VEPs) to be databases or software packages that predict the functional impacts of genetic variants. Each VEP is usually specialized in annotating one or a few categories of variants, such as single nucleotide variations (SNVs), indels, missense variants, or structural variants (SVs) (Geoffroy et al. 2018; Pagel et al. 2019; Rentzsch et al. 2019; Vaser et al. 2016). The variety of VEPs and their functionalities poses the challenge of choosing the appropriate tool for a specific task, a topic that has been addressed in non-systematic reviews

✉ Andreas Ziegler
ziegler.lit@mailbox.org

1 Cardio-CARE, Medizincampus Davos, Herman-Burchard-Str. 1, Davos Wolfgang, 7265 Davos, Switzerland

2 Swiss Institute of Bioinformatics, Lausanne, Switzerland

3 Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

4 University Center of Cardiovascular Science & Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

5 School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

(Katsonis et al. 2022; Tabarini et al. 2022). Some reviews summarize VEPs for one type of variant only (Abramowicz and Gos 2018; Glusman et al. 2017). Other articles focus on variation relevant to the American College of Medical Genetics and Genomics/Association of Molecular Pathology (ACMG/AMP) guidelines (Ghosh et al. 2017; Kassahn et al. 2014). All reviews have in common that their summary tables group functional information into a few categories, usually SNVs, indels, and SVs only. This categorization limits the search for VEPs suitable for other categories of variants, such as missense mutations or copy number variation.

This work aims to provide a systematic overview of the broad range of variant types and their functional impacts across VEPs. To this end, we systematically searched MEDLINE and investigated the possible input and output of each tool. The efficient selection of the most appropriate tool for a specific task can easily be accomplished using an interactive website.

Methods

A systematic review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al. 2021). The protocol was registered in OSF Registries on November 10, 2023 (https://doi.org/10.17605/OSF.IO/S2GCT).

Literature search

The literature search was conducted in the MEDLINE database. The search was restricted to articles published in English after January 1, 2014. This date was chosen to coincide with two milestones in genomics: the launch of the GRCh38 reference genome in December 2013 and the release of higher-throughput sequencing machines (Guo et al. 2017; Sheridan 2014). The search was performed on November 10, 2023, and results up to that date were included. The search query combined groups of terms related to variant, effect, prediction, and tools. Within each group, the terms were combined using the logical operator OR. The complete query is provided in Supplementary Table S1.

Articles containing the term "cancer" in the title were excluded to reduce the number of irrelevant hits and to find VEPs applicable across several diseases. We scanned the reference lists of review and benchmarking articles to retrieve additional eligible articles.

Study selection

Included articles described a VEP, i.e., a tool accepting human genetic variants and predicting functional impacts. The list of exclusion criteria was made to ensure that tools were reliable, broadly applicable, accessible, scalable, and reproducible (Table 1). In cases where a tool appeared to be discontinued, generally indicated by a non-functional URL in the publication, we contacted the corresponding author for confirmation. Some authors supplied a working URL, which allowed us to reassess the publication against the other exclusion criteria. We removed tools not applicable to humans or without any documentation.

Review and benchmarking articles were used to find additional eligible articles. However, only original work describing a VEP was included in this review. Web-only and GUI-only tools were deemed insufficiently reproducible and scalable and were thus excluded. In line with our accessibility requirement, tools requiring a fee were also excluded. Additionally, given the fast pace of progress in the field, we included only tools that support the GRCh38 genome build and were updated at least once since January 1, 2020. Tools that were specific to a small number of genes or a specific disease were excluded, as we were interested in the application of VEPs to a broad range of studies. If several versions of the tool existed, we only included the latest version, regardless of whether the latest version had an associated publication. Nevertheless, significant updates often coincided with a publication, such as dbNSFP v4 (Liu et al. 2020).

One author (CR) selected the studies based on the exclusion criteria (Table 1). First, titles were screened for eligibility. Second, articles were filtered based on the abstract. Third, the full text of the remaining articles was examined. Reasons for exclusion were recorded for each round.

Data extraction

First, one author (CR) extracted the tool name, variant types, functional impacts, and operating system requirements from

Table 1 Exclusion criteria

Not a database, tool, or score
Discontinued tool
Newer version available
Not applicable to humans
Not a VEP
No documentation
Preprint
Review or benchmarking publications
Not easily downloadable, e.g., web-only or GUI-only tool
Not completely free
Not supporting the GRCh38 genome build
Specific to a small number of genes
Specific to a disease
Not updated since January 1, 2020

the included publications and their latest documentation. The URLs of tools with online capabilities were retrieved. Tools that required a high-performance computer were identified. Second, another author (LG) reviewed the extracted data to confirm the accuracy of the information from the publications and documentation. Divergences were resolved through discussion. For each article, the following characteristics were automatically retrieved: PubMed ID, title, authors, citation, first author, journal, year of publication, date of PubMed entry creation, PMCID, NIHMS ID, and digital object identifier.

Sequence Ontology terms were used to describe the variant types and functional impacts wherever applicable (Eilbeck et al. 2005). In case a Sequence Ontology term was unavailable to describe a particular variant type or functional impact, a new term was coined. For terms consistent with the structure of the Sequence Ontology, a request to create the new term was made on the Sequence Ontology GitHub page (https://github.com/The-Sequence-Ontology/SO-Ontologies/issues). Eighteen new terms were requested and are awaiting approval. Examples include "enhancer variant" and "promoter variant". The full list of Sequence Ontology terms is provided in Supplementary Table S2.

Data synthesis

Descriptive statistics were calculated for each tool, including the number of variant and functional impact categories. Linear regression was used to study the relationship between the number of functional impacts predicted by each tool and the date it was uploaded to the MEDLINE database.

Software

All analyses used R version 4.2.2; all R scripts are attached as supplementary files and were uploaded to Zenodo (see section Code availability). A website was created with the shiny package (Chang et al. 2023).

Results

The MEDLINE query yielded 7273 records, of which 6514 were excluded after title screening (Fig. 1). Abstract screening excluded an additional 542 records, leaving 217 full-text articles for eligibility assessment. Detailed reading led to the exclusion of 120 articles. The most frequent reasons for exclusion were the fact that the work did not describe a VEP and the lack of maintenance. Examining references from benchmarking and review articles added 21 relevant publications. In total, this review encompasses 118 original articles, each covering a unique VEP, and references to all 118 articles are provided in Supplementary Table S3

Overview of VEPs

The 118 VEPs differed in both the accepted variant types and the predicted functional impacts (Fig. 2). The number of accepted variant types per VEP ranged from one to seven. Seventy-three VEPs specialized in a single variant type, and two tools, Ensembl VEP and DECIPHER, accepted seven types (Bragin et al. 2014; McLaren et al. 2016). The number of predicted functional impacts per tool ranged from 1 to 58, and approximately two thirds (n = 82) predicted a single functional impact. SnpEff stood out as the VEP with the most predicted functional impacts (Cingolani et al. 2012). Some databases achieved many predicted impacts by aggregating predictions from multiple sources. For example, FAVOR and WGSA aggregated 48 and 40 annotations, respectively (Liu et al. 2016; Zhou et al. 2023).

The implementation of VEPs varied substantially. For example, AbraOM was a simple text file that lists variants and their functional impacts, while SparkINFERNO was a complex software package requiring a high-performance computing environment for execution (Kuksa et al. 2020; Naslavsky et al. 2017).

Variant types

A total of 36 distinct variant types were accepted by VEPs, such as SNVs, indels, CNVs (Supplementary Table S4, Ding et al. 2023). Other acceptable inputs included variants falling in specific regions, such as splice sites, introns, exons, untranslated regions (UTRs), promoters, or enhancers. Despite the diverse range of variants being accepted across tools, some clinically and biologically important variants were missing. Gene fusions were unsupported by VEPs, both as input and predicted functional impacts. Furthermore, variant types were not equally represented across the tools (Fig. 3). While 69 VEPs accepted SNVs as input, 19 other variant types were accepted by only one VEP. For example, DECIPHER was the only database containing inversions and translocations (Bragin et al. 2014).

Functional impacts

VEPs predicted 161 distinct functional impacts (Supplementary Table S5). Pathogenicity was the most common functional impact, predicted by 75 VEPs, because of its clinical relevance and use within ACMG/AMP guidelines (Fig. 3; Richards et al. 2015). Variant frequency, stop gain, and missense variant were the next most commonly predicted functional impacts (Fig. 3). Although variant frequency technically is not a functional impact, it has been reported by some databases and can provide insight into the evolutionary context and potential benignity of certain variants. Functional impacts are generally classified into

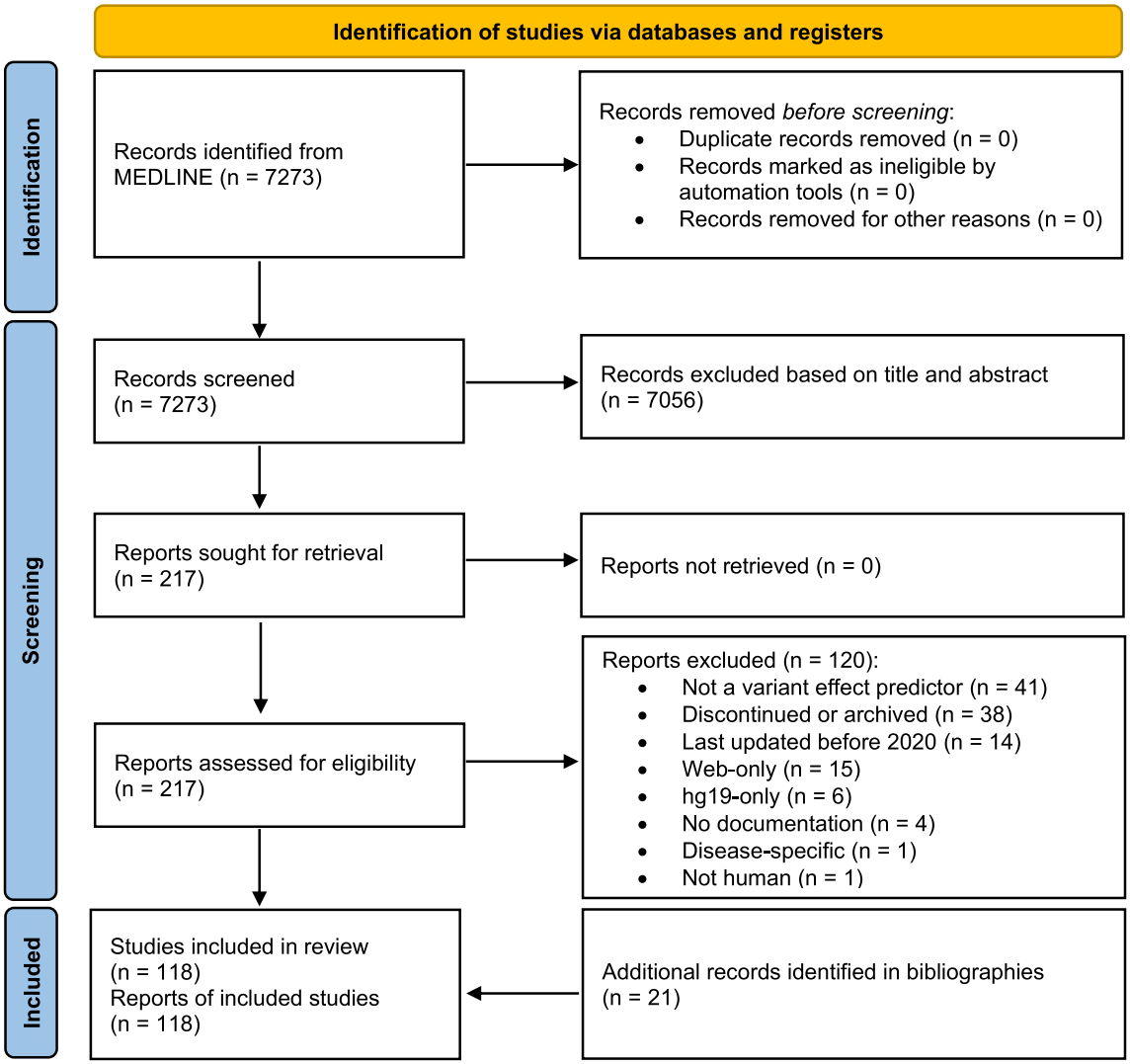


Fig. 1 PRISMA flow diagram of the literature search and selection process

several categories, and effects on the protein sequence, such as missense variants and frameshift indels, formed one well-studied class. Effects on splicing and regulatory elements, e.g., transcription factor-binding sites and enhancers, formed additional categories. Some tools predicted functional impacts not supported by other tools. For example, UTRannotator was the sole predictor of five specific changes in the open reading frames of 5′ UTRs (Zhang et al. 2021). Eighty-nine functional impacts were predicted by only one tool, making these tools indispensable for a study aiming to interpret those impacts.

Only Ensembl VEP, SnpEff, and VAGrENT used a controlled vocabulary, the Sequence Ontology, to describe functional impacts (Cingolani et al. 2012; McLaren et al. 2016; Menzies et al. 2015). However, none of the tools used the Sequence Ontology to describe variant types. Controlled vocabularies may also be used for phenotypes, such as the Experimental Factor Ontology in the GWAS Catalog or Human Phenotype Ontology for DECIPHER.

Figure 4 displays the number of functional impacts by date of publication. While two aggregators (FAVOR and Ensembl VEP) had a publication in the last two years, SnpEff and WGSA were published more than seven years ago. The slope of the linear regression line was –1.03 (95% confidence interval –1.68, –0.37) functional impacts/year, which was significantly different from 0 ($p = 0.002$). As the assumption of linearity in this regression model is questionable, we also ran a quantile regression for the median. The quantile regression revealed a slope of 0 (95% confidence interval 0, 0) and an intercept of 1, indicating no change in the median number of predicted functional impacts over time. This result is due to the 82 VEPs that predict only one functional impact. The number of functional impacts presented in Fig. 4 does not necessarily correspond to the

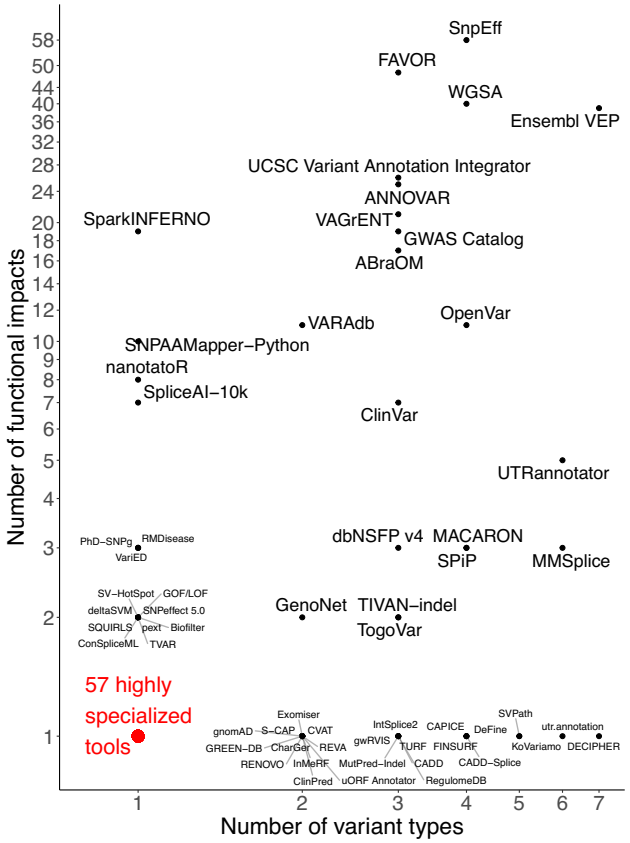


Fig. 2 Scatter plot of the 118 variant effect predictors (VEPs). The number of functional impacts is plotted against the number of accepted variant types. To avoid overplotting, tools that annotate a single variant type and output only one functional impact type are collectively represented by a single red dot

number predicted in that year. This discrepancy can occur because tools are updated, and we extracted data from the most recent documentation.

Supplementary Fig. S1 shows the number of variant types supported for the first time each year. There was no statistically significant upward or downward trend ($p = 0.212$).

A Shiny app to find VEPs

To identify suitable VEPs, a Shiny app website can be used. The website features a searchable table listing the tools along with the specific variant types and functional impacts they address. Users can filter this table using Sequence Ontology terms relevant to their needs. They can also enter the number of VEPs they wish to implement. The site will then display the tools that maximize the number of impact predictions. For example, the top three tools, SnpEff, FAVOR, and SparkInferno, predict 99 different impacts, thus cover 61% of all possible impacts. In this combination, SnpEff covers 54 functional impacts covering changes in the coding sequence, UTRs, gene structure, regulatory regions, splicing, and others. FAVOR adds 36 annotations related to histone modifications, pathogenicity scores, and disease associations. SparkINFERNO adds 9 annotations related to non-coding RNAs and regulatory regions. Users may also filter VEPs according to the supported operating system and the availability of an online version of the tool. The website also includes a bibliography of review and benchmark articles. Benchmark studies that compare the performance of various tools provide valuable assistance in refining tool selection. One study's results, including sensitivity, specificity, positive predictive value, and negative predictive value, are accessible in a searchable table. REVEL is the best performer according to all four metrics (Ghosh et al. 2017). A

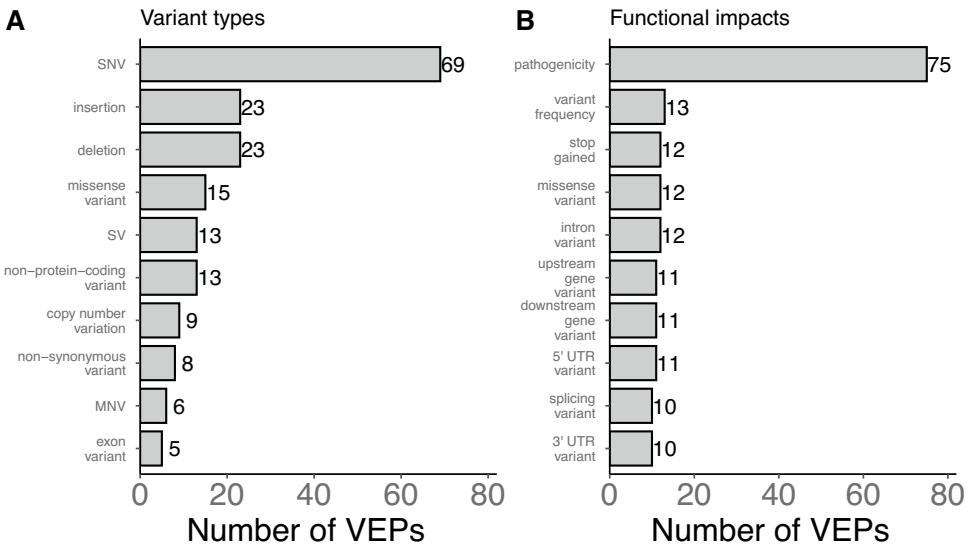


Fig. 3 Barplots of the most commonly annotated variant types and predicted functional impacts. **A** Bars represent the number of tools accepting each of the 10 most commonly accepted variant types. **B** Bars represent the number of tools predicting each of the 10 most commonly predicted functional impacts

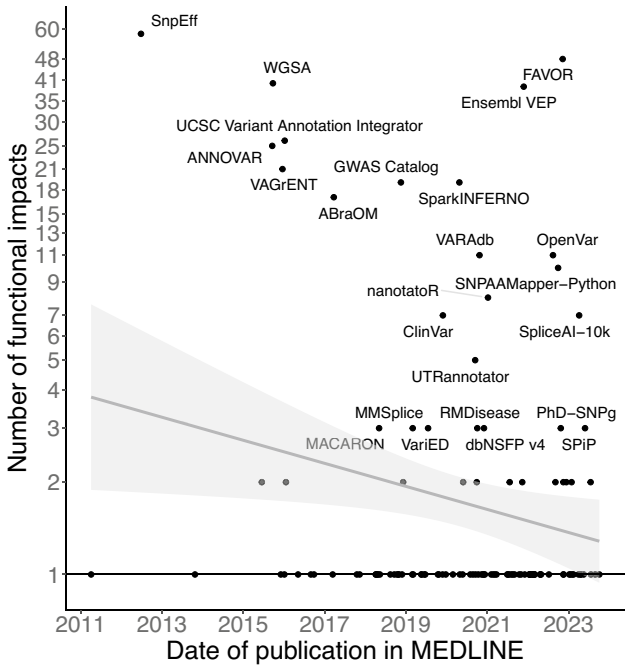


Fig. 4 Number of predicted functional impacts annotated by each tool over time. Tools with ≥3 predictions are labeled. The x-axis shows the day of publication in MEDLINE. The y-axis shows the number of predicted functional impacts in the most recent version of the tool. The grey line represents the least squares regression line. The light grey shaded area surrounding this line represents the 95% pointwise confidence intervals. The flat black line represents the quantile regression line at the median

recent benchmark study identified ClinPred and REVEL as the top pathogenicity predictors in this review, with the ranking of all 55 evaluated predictors accessible via the Shiny app (Livesey and Marsh 2023). While some tools performed better than ClinPred and REVEL, they were excluded from our review for not meeting the inclusion criteria. For example, ESM-1v was not peer-reviewed (Meier et al. 2021), and EVE and DeepSequence did not accept genetic variants (Frazer et al. 2021; Riesselman et al. 2018).

Discussion

This systematic review identified 118 VEPs that together accepted 36 variant types and predicted 161 functional impacts. The functionalities of these numerous VEPs exhibited considerable diversity. Some VEPs accepted only one variant type, while others accepted up to 7. Similarly, some VEPs predicted a single functional impact, while others predicted up to 58. About half of these tools were highly specialized and predicted a single functional impact for one variant type. In contrast, SnpEff, FAVOR, and SparkINFERNO, could predict more than 40 functional impacts each. Using only these three VEPs covered 61% of the predictable functional impacts. Additionally, 75 tools predicted pathogenicity, making them usable as supporting diagnostic evidence according to the ACMG/AMP guidelines (Richards et al. 2015). To facilitate the selection of VEPs, we launched an interactive website that presents a list of tools according to user needs.

Out of 217 full-text articles analyzed, 38 VEPs have been completely discontinued, and another 14 have not been updated since 2019. The lack of maintenance of biological databases and tools is a recurring issue (Imker 2018), which causes difficulty for researchers who depend on these resources. Financial constraints and limited value may justify the discontinuation of a database. However, the database should be archived in a repository to ensure reproducibility (Imker 2018).

Our search was limited to MEDLINE and English language articles, potentially missing relevant studies in other databases or languages. To mitigate these limitations, we examined the references in review and benchmarking articles found by our search to find missed publications. Furthermore, we note that two recent reviews on rare non-coding variant annotation tools encompass 40 and 30 tools, respectively (Kuksa et al. 2022; Tabarini et al. 2022). One benchmarking paper covers 55 VEPs (Livesey and Marsh 2023). Thus, our review stands as the most comprehensive with 118 VEPs described.

In the screening process, we removed tools that were disease-specific, gene-specific, web-only, not updated since 2019, or rely on the hg19 genome build. While some tools may still be useful in specific contexts, we excluded them from our review to focus on broadly applicable, up-to-date, and scalable tools. While web-only databases are easily accessible, they lack the reproducibility, scalability, and privacy of downloadable VEPs.

Another limitation of our review is that it does not provide the total number of annotated variants for each tool. For example, while CADD scored each of the possible ~ 9 billion human SNVs, the ABraOM database contained only 2.3 million variants. In fact, it is difficult to provide the precise number of annotated variants for each tool, as databases are subject to constant updates. Any number reported in the original publication is most likely outdated.

Adoption of a common standardized vocabulary would improve the comparison, integration, and discoverability of VEPs (Brookes and Robinson 2015). Only Ensembl VEP, SnpEff, and VAGrENT used a controlled vocabulary to describe functional impacts (Cingolani et al. 2012; McLaren et al. 2016; Menzies et al. 2015). In this review, we standardized the input and output terms used by each tool according to the Sequence Ontology (Eilbeck et al. 2005). This approach facilitates the search for tools in VEP Finder. For example, users interested in VEPs that accept 5′ UTR

variants can choose '5_prime_UTR_variant' from a drop-down menu, thereby avoiding confusion over non-standard terms. VEP Finder will then display all the tools that accept variants in 5' UTRs. The term 'pathogenicity' is widely used across the 75 pathogenicity predictors to describe variant impact, usually on a scale from 'benign' to 'pathogenic.' However, the inconsistent scale and vocabulary across tools, with some using terms like 'neutral,' 'tolerated,' or 'deleterious,' complicates direct comparisons. While most tools focus on disease relevance, some, such as SIFT, assess the effect on protein function (Vaser et al. 2016). The ACMG/AMP guidelines provide a standardized framework for defining pathogenicity concerning disease, but no similar classification guidelines exist for protein function. Thus, authors must clearly define what they mean by "pathogenicity" and how to interpret the scores.

Selecting suitable VEPs requires considering parameters beyond the accepted input and predicted output. Metrics, such as accuracy and precision, help in identifying tools with higher analytical performance (Livesey and Marsh 2023; Pejaver et al. 2022). Once the selection of VEPs has been made, guidance exists to interpret their outputs (Cheng et al. 2020). For diagnostic purposes, clinicians are advised to consult the ACMG/AMP guidelines to use VEPs (Richards et al. 2015), and for a limited number of genes, more detailed guidance on variant interpretation is available (Fortuno et al. 2021; Lee et al. 2018).

Future research

This review revealed that no VEP accepts gene fusions as input. This gap may be due to their lower frequency in the human population and because of the limitations of second-generation sequencing technologies. However, their clinical importance calls for support soon (Nelson et al. 2017). New variant types were regularly supported (Fig. S1). Should this trend continue, more variant types will likely receive support in the coming years.

The absence of a benchmarking study assessing all 75 pathogenicity predictors highlights the difficulty of this endeavor. A meta-analysis of existing studies could shed light on the best-performing VEPs and might discriminate between the many pathogenicity predictors. This analysis would need to account for the variability in the sets of tools and testing datasets used across different studies.

To maximize the utility of VEPs for clinical and research purposes, further advancements are required to extend predictions specific to isoforms, tissues, and traits to more variants. Such functionality will enhance our understanding of variant effects and facilitate their experimental validation. Moreover, developing trait-specific pathogenicity scores is essential because certain variants may be pathogenic for one disease but benign or even advantageous for another (Taylor et al. 2012). Furthermore, to facilitate interoperability between different tools, we also advocate the use of controlled vocabularies to describe phenotypes (Kohler et al. 2021; Malone et al. 2010). We aim to perform a bigger update of the VEP Finder once per year and to do regular update after user input and evidence.

VEPs predicting many functional impacts, such as SnpEff, FAVOR and WGSA, represent a potential solution to the problem of tool choice. Nevertheless, the rapid evolution of the field necessitates continuous updates to keep them up to date. Furthermore, we expect specialized tools to be continuously released (Fig. 4). Consequently, systematic reviews on VEPs will be needed regularly.

Conclusion

A staggering 118 tools were available to predict approximately 160 functional impacts that ranged from molecular to phenotypic effects. About 60% of these impacts could be predicted by combining just three tools. Unexpectedly, recent tools did not necessarily predict more impacts than older ones. Despite the vast diversity of VEPs, some genetic variants were not yet supported and should be the object of future research.

The abundance of available options can complicate the tool selection process. However, this challenge is mitigated by the Shiny app developed in this review. The app enables users to filter tools based on their specific needs, narrowing down the list of suitable options.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-024-02670-5.

Acknowledgements We thank Dr. Hugo Solleder for his feedback on the research questions and the VEP Finder website.

Author contributions Design, conception of the study and development of search strategy: Cristian Riccio, Max L. Jansen, Andreas Ziegler. Literature search: Cristian Riccio. Data collection: Cristian Riccio and Linlin Guo. Data analysis and interpretation: Cristian Riccio, Andreas Ziegler. Drafted manuscript: Cristian Riccio, Andreas Ziegler. Critical review and approval of final manuscript: all authors.

Availability of data and material The VEP Finder website is freely available at https://cardio-care.shinyapps.io/VEP_Finder/.

Code availability The scripts for data wrangling, generating plots, and running linear models are stored on Zenodo and can be accessed at https://doi.org/10.5281/zenodo.10255446. They are also available in the Supplementary zip file. The code for the Shiny app is accessible at https://github.com/CristianRiccio/VEP_Finder.

Declarations

Conflict of interest CR, MLJ, and AZ are employees of Cardio-CARE AG, a not-for-profit company financed by the Kühne Foundation. AZ is a member of the editorial board of Human Genetics.

Ethics approval Not Applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abramowicz A, Gos M (2018) Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet 59:253–268. https://doi.org/10.1007/s13353-018-0444-7

Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ (2014) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res 42:D993–D1000. https://doi.org/10.1093/nar/gkt937

Brookes AJ, Robinson PN (2015) Human genotype–phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16:702–715. https://doi.org/10.1038/nrg3932

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023) shiny: Web application framework for R. https://CRAN.R-project.org/package=shiny. Accessed 3 Apr 2024

Cheng N, Li M, Zhao L, Zhang B, Yang Y, Zheng CH, Xia J (2020) Comparison and integration of computational methods for deleterious synonymous mutation prediction. Brief Bioinform 21:970–981. https://doi.org/10.1093/bib/bbz047

Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (austin) 6:80–92. https://doi.org/10.4161/fly.19695

Ding Q, Somerville C, Manshaei R, Trost B, Reuter MS, Kalbfleisch K, Stanley K, Okello JBA, Hosseini SM, Liston E, Curtis M, Zarrei M, Higginbotham EJ, Chan AJS, Engchuan W, Thiruvahindrapuram B, Scherer SW, Kim RH, Jobling RK (2023) SCIP: software for efficient clinical interpretation of copy number variants detected by whole-genome sequencing. Hum Genet 142:201–216. https://doi.org/10.1007/s00439-022-02494-1

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The Sequence ontology: a tool for the unification of genome annotations. Genome Biol 6:R44. https://doi.org/10.1186/gb-2005-6-5-r44

Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS (2021) Disease variant prediction with deep generative models of evolutionary data. Nature 599:91–95. https://doi.org/10.1038/s41586-021-04043-8

Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J (2018) AnnotSV: an integrated tool for structural variations annotation. Bioinform 34:3572–3574. https://doi.org/10.1093/bioinformatics/bty304

Ghosh R, Oak N, Plon SE (2017) Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol 18:225. https://doi.org/10.1186/s13059-017-1353-5

Glusman G, Rose PW, Prlic A, Dougherty J, Duarte JM, Hoffman AS, Barton GJ, Bendixen E, Bergquist T, Bock C, Brunk E, Buljan M, Burley SK, Cai B, Carter H, Gao J, Godzik A, Heuer M, Hicks M, Hrabe T, Karchin R, Leman JK, Lane L, Masica DL, Mooney SD, Moult J, Omenn GS, Pearl F, Pejaver V, Reynolds SM, Rokem A, Schwede T, Song S, Tilgner H, Valasatava Y, Zhang Y, Deutsch EW (2017) Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med 9:113. https://doi.org/10.1186/s13073-017-0509-y

Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y (2017) Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics 109:83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, Palsson G, Hardarson MT, Oddsson A, Jensson BO, Kristmundsdottir S, Sigurpalsdottir BD, Stefansson OA, Beyter D, Holley G, Tragante V, Gylfason A, Olason PI, Zink F, Asgeirsdottir M, Sverrisson ST, Sigurdsson B, Gudjonsson SA, Sigurdsson GT, Halldorsson GH, Sveinbjornsson G, Norland K, Styrkarsdottir U, Magnusdottir DN, Snorradottir S, Kristinsson K, Sobech E, Jonsson H, Geirsson AJ, Olafsson I, Jonsson P, Pedersen OB, Erikstrup C, Brunak S, Ostrowski SR, Consortium DG, Thorleifsson G, Jonsson F, Melsted P, Jonsdottir I, Rafnar T, Holm H, Stefansson H, Saemundsdottir J, Gudbjartsson DF, Magnusson OT, Masson G, Thorsteinsdottir U, Helgason A, Jonsson H, Sulem P, Stefansson K (2022) The sequences of 150,119 genomes in the UK Biobank. Nature 607:732–740. https://doi.org/10.1038/s41586-022-04965-x

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature 470:264–268. https://doi.org/10.1038/nature09753

Imker HJ (2018) 25 years of molecular biology databases: a study of proliferation, impact, and maintenance. Front Res Metr Anal 3:18. https://doi.org/10.3389/frma.2018.00018

Kassahn KS, Scott HS, Caramins MC (2014) Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. Hum Mutat 35:413–423. https://doi.org/10.1002/humu.22525

Katsonis P, Wilhelm K, Williams A, Lichtarge O (2022) Genome interpretation using in silico predictors of variant impact. Hum Genet 141:1549–1577. https://doi.org/10.1007/s00439-022-02457-6

Kohler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yuksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN (2021) The human phenotype ontology in 2021. Nucleic Acids Res 49:D1207–D1217. https://doi.org/10.1093/nar/gkaa1043

Kuksa PP, Lee CY, Amlie-Wolf A, Gangadharan P, Mlynarski EE, Chou YF, Lin HJ, Issen H, Greenfest-Allen E, Valladares O, Leung YY, Wang LS (2020) SparkINFERNO: a scalable high-throughput pipeline for inferring molecular mechanisms of noncoding genetic variants. Bioinformatics 36:3879–3881. https://doi.org/10.1093/bioinformatics/btaa246

Kuksa PP, Greenfest-Allen E, Cifello J, Ionita M, Wang H, Nicaretta H, Cheng PL, Lee WP, Wang LS, Leung YY (2022) Scalable approaches for functional analyses of whole-genome sequencing non-coding variants. Hum Mol Genet 31:R62–R72. https://doi.org/10.1093/hmg/ddac191

Lee K, Krempely K, Roberts ME, Anderson MJ, Carneiro F, Chao E, Dixon K, Figueiredo J, Ghosh R, Huntsman D, Kaurah P, Kesserwan C, Landrith T, Li S, Mensenkamp AR, Oliveira C, Pardo C, Pesaran T, Richardson M, Slavin TP, Spurdle AB, Trapp M, Witkowski L, Yi CS, Zhang L, Plon SE, Schrader KA, Karam R (2018) Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. Hum Mutat 39:1553–1568. https://doi.org/10.1002/humu.23650

Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH, Huang Z, Carroll A, Wei P, Gibbs R, Klein RJ, Boerwinkle E (2016) WGSA: an annotation pipeline for human genome sequencing studies. J Med Genet 53:111–112. https://doi.org/10.1136/jmedgenet-2015-103423

Liu X, Li C, Mou C, Dong Y, Tu Y (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 12:103. https://doi.org/10.1186/s13073-020-00803-9

Livesey BJ, Marsh JA (2023) Updated benchmarking of variant effect predictors using deep mutational scanning. Mol Syst Biol 19:e11474. https://doi.org/10.15252/msb.202211474

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H (2010) Modeling sample variables with an experimental factor ontology. Bioinformatics 26:1112–1118. https://doi.org/10.1093/bioinformatics/btq099

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F (2016) The Ensembl variant effect predictor. Genome Biol 17:122. https://doi.org/10.1186/s13059-016-0974-4

Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. Advances Neural Inf Process Syst. https://doi.org/10.1101/2021.07.09.450648

Menzies A, Teague JW, Butler AP, Davies H, Tarpey P, Nik-Zainal S, Campbell PJ (2015) VAGrENT: variation annotation generator. Curr Protoc Bioinform 52:1. https://doi.org/10.1002/0471250953.bi1508s52

Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, Bozoklian D, Sandberg TOM, Brito LA, Lazar M, Bernardo DV, Amaro E Jr, Duarte YAO, Lebrao ML, Passos-Bueno MR, Zatz M (2017) Exomic variants of an elderly cohort of Brazilians in the ABraOM database. Hum Mutat 38:751–763. https://doi.org/10.1002/humu.23220

Nelson KN, Peiris MN, Meyer AN, Siari A, Donoghue DJ (2017) Receptor tyrosine kinases: translocation partners in hematopoietic disorders. Trends Mol Med 23:59–79. https://doi.org/10.1016/j.molmed.2016.11.002

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hrobjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

Pagel KA, Antaki D, Lian A, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P (2019) Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. PLoS Comput Biol 15:e1007112. https://doi.org/10.1371/journal.pcbi.1007112

Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS, Biesecker LG, Radivojac P, Brenner SE, ClinGen Sequence Variant Interpretation Working G (2022) Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet 109:2163–2177. https://doi.org/10.1016/j.ajhg.2022.10.013

Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB (2019) Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat Rev Genet 20:747–759. https://doi.org/10.1038/s41576-019-0177-4

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acid Res 47:D886–D894. https://doi.org/10.1093/nar/gky1016

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee ALQA (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–424. https://doi.org/10.1038/gim.2015.30

Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture the effects of mutations. Nat Methods 15:816–822. https://doi.org/10.1038/s41592-018-0138-4

Sheridan C (2014) Milestone approval lifts Illumina's NGS from research into clinic. Nat Biotechnol 32:111–112. https://doi.org/10.1038/nbt0214-111

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Gunes O, Hall P, Hayhurst J, Ibrahim A, Ji Y, John S, Lewis E, MacArthur JAL, McMahon A, Osumi-Sutherland D, Panoutsopoulou K, Pendlington Z, Ramachandran S, Stefancsik R, Stewart J, Whetzel P, Wilson R, Hindorff L, Cunningham F, Lambert SA, Inouye M, Parkinson H, Harris LW (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acid Res 51:D977–D985. https://doi.org/10.1093/nar/gkac1010

Tabarini N, Biagi E, Uva P, Iovino E, Pippucci T, Seri M, Cavalli A, Ceccherini I, Rusmini M, Viti F (2022) Exploration of tools for the interpretation of human non-coding variants. Int J Mol Sci. https://doi.org/10.3390/ijms232112977

Taub MA, Conomos MP, Keener R, Iyer KR, Weinstock JS, Yanek LR, Lane J, Miller-Fleming TW, Brody JA, Raffield LM, McHugh CP, Jain D, Gogarten SM, Laurie CA, Keramati A, Arvanitis M, Smith AV, Heavner B, Barwick L, Becker LC, Bis JC, Blangero J, Bleecker ER, Burchard EG, Celedon JC, Chang YPC, Custer B, Darbar D, de Las FL, DeMeo DL, Freedman BI, Garrett ME, Gladwin MT, Heckbert SR, Hidalgo BA, Irvin MR, Islam T, Johnson WC, Kaab S, Launer L, Lee J, Liu S, Moscati A, North KE, Peyser PA, Rafaels N, Seidman C, Weeks DE, Wen F, Wheeler MM, Williams LK, Yang IV, Zhao W, Aslibekyan S, Auer PL, Bowden DW, Cade BE, Chen Z, Cho MH, Cupples LA, Curran JE, Daya M, Deka R, Eng C, Fingerlin TE, Guo X, Hou L, Hwang SJ, Johnsen JM, Kenny EE, Levin AM, Liu C, Minster RL, Naseri T, Nouraie M, Reupena MS, Sabino EC, Smith JA, Smith NL, Su JL, Taylor JG, Telen MJ, Tiwari HK, Tracy RP, White MJ, Zhang Y, Wiggins KL, Weiss ST, Vasan RS, Taylor KD, Sinner MF, Silverman EK, Shoemaker MB, Sheu WH, Sciurba F, Schwartz DA, Rotter JI, Roden D, Redline S, Raby BA et al (2022) Genetic determinants of telomere length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. Cell Genom. https://doi.org/10.1016/j.xgen.2021.100084

Taylor SM, Parobek CM, Fairhurst RM (2012) Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. Lancet Infect Dis 12:457–468. https://doi.org/10.1016/S1473-3099(12)70055-5

The All of Us Research Program Investigators (2019) The "All of Us" research program. N Engl J Med 381:668–676. https://doi.org/10.1056/NEJMsr1809937

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016) SIFT missense predictions for genomes. Nat Protoc 11:1–9. https://doi.org/10.1038/nprot.2015.123

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678. https://doi.org/10.1038/nature05911

Zhang X, Wakeling M, Ware J, Whiffin N (2021) Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics 37 8:1171–1173. https://doi.org/10.1093/bioinformatics/btaa783

Zhou H, Arapoglou T, Li X, Li Z, Zheng X, Moore J, Asok A, Kumar S, Blue EE, Buyske S, Cox N, Felsenfeld A, Gerstein M, Kenny E, Li B, Matise T, Philippakis A, Rehm HL, Sofia HJ, Snyder G, NHGRI Genome Sequencing Program Variant Functional Annotation Working Group, Weng Z, Neale B, Sunyaev SR, Lin X (2023) FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. Nucleic Acids Res 51:D1300–D1311. https://doi.org/10.1093/nar/gkac966

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

D.11 Protein–protein interactions

Further explanations for Chapter 11.

D.11.1 Experimental methods

- Snider et al. (2015) — Explains the principles of interactome mapping, compares binary assays (yeast two-hybrid and relatives) with co-complex methods such as AP-MS, and discusses their strengths, weaknesses, follow-up validation and the main interaction databases. A good companion to the whole section, including errors, biases and coverage. *(review; CC BY 4.0)*

  Fundamentals of protein interaction network mapping. Snider et al. (2015). *Molecular Systems Biology*. doi:10.15252/msb.20156351

  > Studying protein interaction networks of all proteins in an organism ("interactomes") remains one of the major challenges in modern biomedicine. Such information is crucial to understanding cellular pathways and developing effective therapies for the treatment of human diseases. Over the past two decades, diverse biochemical, genetic, and cell biological methods have been developed to map interactomes. In this review, we highlight basic principles of interactome mapping. Specifically, we discuss the strengths and weaknesses of individual assays, how to select a method appropriate for the problem being studied, and provide general guidelines for carrying out the necessary follow-up analyses. In addition, we discuss computational methods to predict, map, and visualize interactomes, and provide a summary of some of the most important interactome resources. We hope that this review serves as both a useful overview of the field and a guide to help more scientists actively employ these powerful approaches in their research.

- Adames et al. (2018) — Reviews how multi-mutant yeast strains are made for genetic interaction screens, comparing classical methods with CRISPR/Cas-based approaches. Useful technical background to the large yeast genetic interaction maps discussed in the subsection on genetic interactions. *(review; CC BY 4.0)*

  Yeast genetic interaction screens in the age of CRISPR/Cas. Adames et al. (2018). *Current Genetics*. doi:10.1007/s00294-018-0887-8

  > The ease of performing both forward and reverse genetics in Saccharomyces cerevisiae, along with its stable haploid state and short generation times, has made this budding yeast the consummate model eukaryote for genetics. The major advantage of using budding yeast for reverse genetics is this organism's highly efficient homology-directed repair, allowing for precise genome editing simply by introducing DNA with homology to the chromosomal target. Although plasmid- and PCR-based genome editing tools are quite efficient, they depend on rare spontaneous DNA breaks near the target sequence. Consequently, they can generate only one genomic edit at a time, and the edit must be associated with a selectable marker. However, CRISPR/Cas technology is efficient enough to permit markerless and multiplexed edits in a single step. These features have made CRISPR/Cas popular for yeast strain engineering in synthetic biology and metabolic engineering applications, but it has not been widely employed for genetic screens. In this review, we critically examine different methods to generate multi-mutant strains in systematic genetic interaction screens and discuss the potential of CRISPR/Cas to supplement or improve on these methods.

- Marku and Pancaldi (2023) — An overview, written for newcomers, of the main types of algorithms for inferring gene regulatory networks from time-series transcriptomics, with guidance on choosing a method. Useful background for the subsection on network inference. *(review; CC BY 4.0)*

  From time-series transcriptomics to gene regulatory networks: A review on inference methods. Marku & Pancaldi (2023). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1011254

> Inference of gene regulatory networks has been an active area of research for around 20 years, leading to the development of sophisticated inference algorithms based on a variety of assumptions and approaches. With the ever increasing demand for more accurate and powerful models, the inference problem remains of broad scientific interest. The abstract representation of biological systems through gene regulatory networks represents a powerful method to study such systems, encoding different amounts and types of information. In this review, we summarize the different types of inference algorithms specifically based on time-series transcriptomics, giving an overview of the main applications of gene regulatory networks in computational biology. This review is intended to give an updated reference of regulatory networks inference tools to biologists and researchers new to the topic and guide them in selecting the appropriate inference method that best fits their questions, aims, and experimental data.

D.11.2 Networks

- Koutrouli et al. (2020) — An introduction to graph theory for biologists: graph types, degree distributions, clustering, motifs, network models, clustering algorithms and tools. Covers the network concepts used throughout the section. *(review; CC BY 4.0)*

  A Guide to Conquer the Biological Network Era Using Graph Theory. Koutrouli et al. (2020). *Frontiers in Bioengineering and Biotechnology*. doi:10.3389/fbioe.2020.00034

  > Networks are one of the most common ways to represent biological systems as complex sets of binary interactions or relations between different bioentities. In this article, we discuss the basic graph theory concepts and the various graph types, as well as the available data structures for storing and reading graphs. In addition, we describe several network properties and we highlight some of the widely used network topological features. We briefly mention the network patterns, motifs and models, and we further comment on the types of biological and biomedical networks along with their corresponding computer- and human-readable file formats. Finally, we discuss a variety of algorithms and metrics for network analyses regarding graph drawing, clustering, visualization, link prediction, perturbation, and network alignment as well as the current state-of-the-art tools. We expect this review to reach a very broad spectrum of readers varying from experts to beginners while encouraging them to enhance the field further.

- Sonawane et al. (2019) — Surveys network types and data sources used in network medicine, and how protein interaction, co-expression and gene regulatory networks are combined with phenotype-specific data to study disease, with example applications. *(review; CC BY 4.0)*

  Network Medicine in the Age of Biomedical Big Data. Sonawane et al. (2019). *Frontiers in Genetics*. doi:10.3389/fgene.2019.00294

  > Network medicine is an emerging area of research dealing with molecular and genetic interactions, network biomarkers of disease, and therapeutic target discovery. Large-scale biomedical data generation offers a unique opportunity to assess the effect and impact of cellular heterogeneity and environmental perturbations on the observed phenotype. Marrying the two, network medicine with biomedical data provides a framework to build meaningful models and extract impactful results at a network level. In this review, we survey existing network types and biomedical data sources. More importantly, we delve into ways in which the network medicine approach, aided by phenotype-specific biomedical data, can be gainfully applied. We provide three paradigms, mainly dealing with three major biological network archetypes: protein-protein interaction, expression-based, and gene regulatory networks. For each of these paradigms, we discuss a broad overview of philosophies under which various network methods work. We also provide a few examples in each paradigm as a test case of its successful application. Finally, we delineate several opportunities and challenges in the field of network medicine. We hope this review provides a lexicon for researchers from biological sciences and network theory to come on the same page to work on research areas that require interdisciplinary expertise. Taken together, the understanding gained from combining biomedical data with networks can be useful for characterizing disease etiologies and identifying therapeutic targets, which, in turn, will lead to better preventive medicine with translational impact on personalized healthcare.

- Gu et al. (2019) — Reviews genome-scale metabolic models, which are simulated to predict metabolic fluxes, and their applications from strain engineering to drug targets and human

disease; background for the subsection on systems biology and metabolic modelling. *(review; CC BY 4.0)*

Current status and applications of genome-scale metabolic models. Gu et al. (2019). *Genome Biology*. doi:10.1186/s13059-019-1730-3

> Genome-scale metabolic models (GEMs) computationally describe gene-protein-reaction associations for entire metabolic genes in an organism, and can be simulated to predict metabolic fluxes for various systems-level metabolic studies. Since the first GEM for Haemophilus influenzae was reported in 1999, advances have been made to develop and simulate GEMs for an increasing number of organisms across bacteria, archaea, and eukarya. Here, we review current reconstructed GEMs and discuss their applications, including strain development for chemicals and materials production, drug targeting in pathogens, prediction of enzyme functions, pan-reactome analysis, modeling interactions among multiple cells or organisms, and understanding human diseases.

D.11.3 Prediction methods

- Tsuchiya et al. (2022) — A short review that traces computational prediction of protein complexes from rigid-body docking and scoring to AlphaFold-based methods, with a historical perspective. Useful for the docking and deep learning subsections. *(review; CC BY 4.0)*

  Protein–protein interaction prediction methods: from docking-based to AI-based approaches. Tsuchiya et al. (2022). *Biophysical Reviews*. doi:10.1007/s12551-022-01032-7

  > Protein-protein interactions (PPIs), such as protein-protein inhibitor, antibody-antigen complex, and supercomplexes play diverse and important roles in cells. Recent advances in structural analysis methods, including cryo-EM, for the determination of protein complex structures are remarkable. Nevertheless, much room remains for improvement and utilization of computational methods to predict PPIs because of the large number and great diversity of unresolved complex structures. This review introduces a wide array of computational methods, including our own, for estimating PPIs including antibody-antigen interactions, offering both historical and forward-looking perspectives.

- Kiouri et al. (2025) — Surveys machine-learning and deep-learning methods that use protein structure to predict interactions and interface residues, and discusses limitations such as scarce structural data and the choice of negative examples, which relates to the evaluation subsection. *(review; CC BY 4.0)*

  Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning. Kiouri et al. (2025). *Biomolecules*. doi:10.3390/biom15010141

  > Protein-Protein Interaction (PPI) prediction plays a pivotal role in understanding cellular processes and uncovering molecular mechanisms underlying health and disease. Structure-based PPI prediction has emerged as a robust alternative to sequence-based methods, offering greater biological accuracy by integrating three-dimensional spatial and biochemical features. This work summarizes the recent advances in computational approaches leveraging protein structure information for PPI prediction, focusing on machine learning (ML) and deep learning (DL) techniques. These methods not only improve predictive accuracy but also provide insights into functional sites, such as binding and catalytic residues. However, challenges such as limited high-resolution structural data and the need for effective negative sampling persist. Through the integration of experimental and computational tools, structure-based prediction paves the way for comprehensive proteomic network analysis, holding promise for advancements in drug discovery, biomarker identification, and personalized medicine. Future directions include enhancing scalability and dataset reliability to expand these approaches across diverse proteomes.

- Lin et al. (2025) — Reviews the use of AlphaFold 3 for complex structure prediction, compares it with docking tools such as ClusPro and with AlphaPulldown, and discusses its limitations for large complexes, dynamics and mutation effects. The examples are from plant and fungal research but the discussion is general. *(review; CC BY 4.0)*

  Analysing protein complexes in plant science: insights and limitation with AlphaFold 3. Lin et al. (2025). *Botanical Studies*. doi:10.1186/s40529-025-00462-2

AlphaFold 3 (AF3), an artificial intelligence (AI)-based software for protein complex structure prediction, represents a significant advancement in structural biology. Its flexibility and enhanced scalability have unlocked new applications in various fields, specifically in plant science, including improving crop resilience and predicting the structures of plant-specific proteins involved in stress responses, signalling pathways, and immune responses. Comparisons with existing tools, such as ClusPro and AlphaPulldown, highlight AF3's unique strengths in sequence-based interaction predictions and its greater adaptability to various biomolecular structures. However, limitations persist, including challenges in modelling large complexes, protein dynamics, and structures from underrepresented plant proteins with limited evolutionary data. Additionally, AF3 encounters difficulties in predicting mutation effects on protein interactions and DNA binding, which can be improved with molecular dynamics and experimental validation. This review presents an overview of AF3's advancements, using examples in plant and fungal research, and comparisons with existing tools. It also discusses current limitations and offers perspectives on integrating molecular dynamics and experimental validation to enhance its capabilities.

The following open-access paper(s) are reproduced under their original licences.

Review

OPEN ACCESS

molecular systems biology

Fundamentals of protein interaction network mapping

Jamie Snider[1], Max Kotlyar[2], Punit Saraon[1], Zhong Yao[1], Igor Jurisica[2] & Igor Stagljar[1,*]

Abstract

Studying protein interaction networks of all proteins in an organism ("interactomes") remains one of the major challenges in modern biomedicine. Such information is crucial to understanding cellular pathways and developing effective therapies for the treatment of human diseases. Over the past two decades, diverse biochemical, genetic, and cell biological methods have been developed to map interactomes. In this review, we highlight basic principles of interactome mapping. Specifically, we discuss the strengths and weaknesses of individual assays, how to select a method appropriate for the problem being studied, and provide general guidelines for carrying out the necessary follow-up analyses. In addition, we discuss computational methods to predict, map, and visualize interactomes, and provide a summary of some of the most important interactome resources. We hope that this review serves as both a useful overview of the field and a guide to help more scientists actively employ these powerful approaches in their research.

Keywords bioinformatics; interactome mapping; PPI technologies; protein-protein interactions (PPIs); proteomics

DOI 10.15252/msb.20156351 | Received 5 June 2015 | Revised 25 October 2015 | Accepted 24 November 2015

Mol Syst Biol. (2015) 11: 848

The importance of studying PPIs

As the basic unit of life, cells represent complex biological entities, whose normal function revolves around a delicate interplay between multiple diverse biomolecular systems. Proteins are vital components of these systems, acting as molecular machines, sensors, transporters, and structural elements (among others), with interactions between proteins, hereinafter called protein–protein interactions (PPIs), being key to their function.

Protein–protein interactions are inherently dynamic in nature, adjusting in response to different stimuli and environmental conditions. This provides considerable flexibility in function and allows cells to adapt in a measured way to changing circumstances. Even a subtle dysfunction of PPIs can have major systemic consequences, perturbing interconnected cellular networks and producing disease phenotypes (Barabási *et al*, 2011). Developing in-depth, dynamic PPI maps is therefore critically important in helping us comprehend these complex processes, and identify new proteins and PPIs suitable for therapeutic intervention.

Over the years, we have seen an emergence and growth of a wide range of exciting technologies for the identification and characterization of PPIs. Selecting "the best" technology for a given research application is thus non-trivial. Here, we highlight the strengths and weaknesses of various methodologies, to aid in selecting the appropriate method for the problem at hand. Note that this review does not aim to cover all PPI methods; instead, we focus on newer approaches and earlier methods that remain widely used, and strongly impacted research.

Key considerations

While numerous methods are available for the large-scale study of PPIs, there is no one "perfect" method for all situations, and each has its own strengths and weaknesses. When selecting a suitable method to study interacting partners of a protein of interest, the following factors should be considered:

1) *The Goal of the Study* must be clearly defined. Discovery-driven studies usually aim to explore interactomes in an unbiased manner on a proteome-wide scale. In contrast, targeted interactome studies focus on a subset of PPIs and therefore confine themselves to smaller libraries or arrays corresponding to a defined set of candidate interaction partners. Different methods are better suited to certain classes of proteins as well as to formats and scales, and selection of one that best matches the research goals is critical.
2) *The Distinct Nature of the PPIs Being Studied.* All PPIs have intrinsic biophysical properties, giving each its own unique features. Some important characteristics to consider are the PPI "strength" (binding affinity), and whether the interaction is transient or stable (Perkins *et al*, 2010). Different bioassays display variable sensitivity, and although generally all can detect stable PPIs, only a fraction are capable of detecting transient interactions. It is also important to determine whether or not posttranslational modifications, co-factors, or

1 Donnelly Centre, Department of Biochemistry, Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
2 Princess Margaret Cancer Center, IBM Life Sciences Discovery Centre, University Health Network, Ontario, Canada
*Corresponding author. Tel: +1 416 946 7828; Fax: +1 416 978 8287; Email: igor.stagljar@utoronto.ca

© 2015 The Authors. Published under the terms of the CC BY 4.0 license

additional binding partners are required (e.g., a PPI may be mediated indirectly through a protein complex), as well as where in the cell interactions are expected to occur, since the selected assay must be compatible with these elements.

3) *Time/Cost Constraints.* Not all methods scale-up equally, and some, while offering powerful advantages on a smaller-scale, can become significantly more expensive and time-consuming as the number of interactions studied increases. Additionally, the time and cost required to develop the necessary reagents (e.g., specific constructs, libraries) needs to be considered.
4) *Specialized Equipment and Expertise.* Finally, it is important to ensure that all necessary resources and knowledge required to fully take advantage of a particular method are available. Although the majority of methods are straightforward, some do require specific instrumentation and expertise. Most methods, especially those that attempt to study interactomes on a genomewide scale, also require strong bioinformatics support for analysis and data cleaning.

Guide to available methods

While many PPI assays exist, we present below some of the newer and more widely used approaches, providing a concise overview of their key principles, advantages, and limitations. Key references for each technique, including examples of their large-scale application, can also be found in Table 1.

The yeast two hybrid (Y2H)

Principle Originally developed 25 years ago (Fields & Song, 1989), the Y2H assay (Fig 1A) remains one of the most popular PPI methods. Y2H-based systems can be used to detect interactions between two proteins, protein and nucleic acid, and also in small-molecule screens (Hamdi & Colas, 2012; Ferro & Trabalzini, 2013). The classic Y2H involves the physical separation of two functional moieties of a transcription factor, specifically a DNA-binding domain (BD) and a transcriptional activation domain (AD), and their fusion to candidate interacting proteins. If a protein bearing an AD interacts with, or comes in close proximity to, a protein bearing a BD, the AD and BD are able to function together as a transcription factor, and direct expression of a reporter gene (Fields & Song, 1989).

Advantages The Y2H approach is simple, well established, and low cost and can be easily set up in most laboratory environments. Y2H is scalable and effective for use in both large-scale screening studies, and smaller efforts investigating specific PPIs. Another benefit is that the assay is carried out *in vivo* in the context of the yeast cell, helping avoid some of the complications and artifacts associated with cell lysis. This assay is best suited for the detection of binary interactions (Hamdi & Colas, 2012; Ferro & Trabalzini, 2013).

Limitations The use of a yeast host means that the PPIs from other organisms may in some cases not be detectable, due to poor expression, or a lack of necessary posttranslational modifications, cofactors, or other binding partners. The method requires that both interacting proteins access the nucleus (in order to drive transcription of reporter), which means that proteins confined to particular cellular environments (e.g., the membrane) cannot be studied in their full-length form. The proteins used in this method are also often overexpressed, which can lead to non-specific interactions. Altogether, these effects can lead to a high false-positive rate, necessitating careful follow-up analysis to identify true, biologically relevant interactions. The readout of this method is also indirect, preventing spatial or temporal analysis of PPIs (Hamdi & Colas, 2012; Ferro & Trabalzini, 2013).

Membrane yeast two hybrid (MYTH)

Principle The MYTH assay (Fig 1B) is designed for the analysis of the interactions of membrane proteins. It is based on a split-ubiquitin approach, whereby the ubiquitin protein is divided into

Table 1. Useful literature references for protein–protein interactions (PPI) methods.

Assay	Relevant literature reviewing or introducing technique	Examples of interaction studies using technique
Y2H	Hamdi and Colas (2012); Ferro and Trabalzini (2013); Stasi *et al* (2015)	Yu *et al* (2008); Weimann *et al* (2013); Rajagopala *et al* (2014); Rolland *et al* (2014); Grossmann *et al* (2015)
MYTH	Snider *et al* (2010); Petschnigg *et al* (2012)	Snider *et al* (2013); Lam *et al* (2015); Gulati *et al* (2015)
LUMIER	Blasche and Koegl (2013)	Barrios-Rodiles *et al* (2005); Xu *et al* (2014); Taipale *et al* (2014); Sahni *et al* (2015)
MAPPIT	Sahni *et al* (2015); Lievens *et al* (2011); Lemmens *et al* (2015)	Lievens *et al* (2009); Bovijn *et al* (2013); Rolland *et al* (2014)
KISS	Lievens *et al* (2014)	Amano *et al* (2015)
BIFC	Kerppola (2008); Zhang *et al* (2015)	Lee *et al* (2011b); Snider *et al* (2013); Cooper *et al* (2015)
MaMTH	Petschnigg *et al* (2014)	–
BRET/FRET	Ciruela (2008); Xie *et al* (2011); Ma *et al* (2014)	Kocan *et al* (2008); Audet *et al* (2010); Mandić *et al* (2014); Sauvageau *et al* (2014)
AP-MS	Dunham *et al* (2012)	Wang and Huang (2008); Babu *et al* (2012); Havugimana *et al* (2012)
BioID-MS	Roux *et al* (2012)	Kim *et al* (2014); Dingar *et al* (2015); Lambert *et al* (2015)
PLA	Koos *et al* (2014)	Chen *et al* (2014)
LRC-TriCEPS	Frei *et al* (2013)	Frei *et al* (2012)
AVEXIS	Sanderson (2008); Kerr and Wright (2012); Sun *et al* (2012)	Bushell *et al* (2008); Martin *et al* (2010); Crosnier *et al* (2011)

 © 2015 The Authors

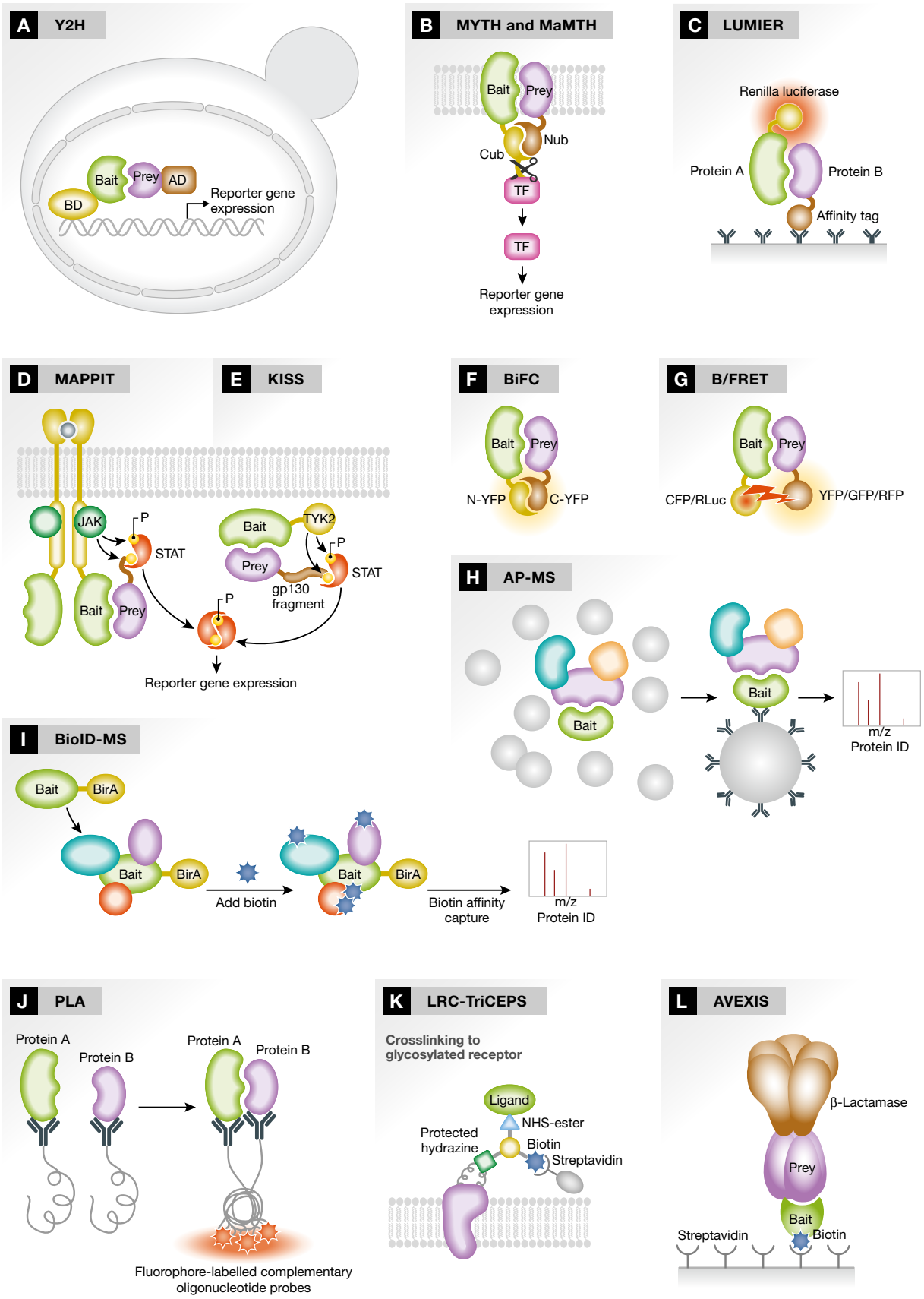


Figure 1.

© 2015 The Authors

Figure 1. Overview of interaction proteomics technologies.
Schematic representations of selected newer and widely used PPI assays. (A) Yeast Two Hybrid (Y2H). (B) Membrane Yeast Two Hybrid (MYTH) and Mammalian Membrane Two Hybrid (MaMTH). (C) Luminescence-based Mammalian Interactome Mapping (LUMIER). (D) Mammalian Protein-Protein Interaction Trap (MAPPIT). (E) Kinase Substrate Sensor (KISS). (F) Bimolecular Fluorescence Complementation (BiFC). (G) Bioluminescence/Fluorescence Resonance Energy Transfer (B/FRET). (H) Affinity Purification-Mass Spectrometry (AP-MS). (I) Proximity-dependent Biotin Identification Coupled to Mass Spectrometry (BioID-MS). (J) Proximity Ligation Assay (PLA). (K) Ligand-Receptor Capture-Trifunctional Chemoproteomics Reagents (LRC-TRiCEPS). (L) Avidity-based Extracellular Interaction Screen (AVEXIS).

two distinct fragments—an N-terminal fragment called "Nub" and a C-terminal fragment called "Cub". The Cub moiety is conjugated to an artificial transcription factor and then fused to a cytosolic terminus of a membrane-bound protein (the "bait"). The Nub moiety is fused to potential interacting partners ("preys"), which can be either membrane-associated or soluble. Interaction of bait and prey proteins brings the Nub and Cub moieties into close proximity, allowing them to form a "pseudoubiquitin" molecule, which is recognized by cellular deubiquitinating enzymes that cleave after the Cub C-terminus. This releases the transcription factor, which then enters the nucleus and activates a reporter system (Stagljar *et al*, 1998; Snider *et al*, 2010).

Advantages Membrane yeast two hybrid is simple, low cost, and scalable for use in both low- and high-throughput (HT) formats. It is easy to establish in any laboratory environment and requires no specialized equipment. The assay is performed *in vivo* in a yeast host, allowing for the study of the interactions of membrane proteins in their full-length form and in the proper context of a membrane environment. This is a significant advantage over the classical Y2H. MYTH is best suited for the detection of binary interactions(Paumi *et al*, 2007; Deribe *et al*, 2009; Snider *et al*, 2010, 2013).

Limitations Membrane yeast two hybrid suffers from some of the same disadvantages as the classical Y2H, including the problems associated with the expression, modification, and interaction of non-native proteins in a yeast host, and artifacts resulting from protein overexpression. Also, MYTH can only be used with membrane proteins that have at least one terminus in the cytosol (where the necessary deubiquitinating enzymes are located). Additionally, soluble proteins cannot be used as baits in the MYTH system, unless they are exceptionally large or anchored to intracellular structures (thereby preventing diffusion of the bait-transcription factor into the nucleus and interaction-independent activation of the reporter system). The readout of this method is also indirect, preventing spatial or temporal analysis of PPIs (Snider *et al*, 2010).

Luminescence-based mammalian interactome mapping (LUMIER)

Principle The LUMIER assay (Fig 1C) is a co-immunoprecipitation-based approach. In this method, one protein ("A") is fused to Renilla luciferase, while another protein ("B") is linked to an affinity tag (e.g., FLAG, HA, protein A). Tagged constructs are transfected into appropriate cell lines where they are overexpressed. Cells are then lysed and protein "B" is immunoprecipitated using an appropriate antibody against the affinity tag. Interaction with protein "A" is assessed by measuring luciferase activity brought down with protein "B" (Barrios-Rodiles *et al*, 2005; Blasche & Koegl, 2013).

Advantages The LUMIER assay is easy to perform and can be used in a HT screening format. It does not require specialized equipment, beyond standard reagents for cell culture and instrumentation to measure bioluminescence. The approach can be used in different cell lines, providing the option of studying PPIs for a given organism in an appropriate *ex vivo* format. Note that this assay is well suited for studying binary interactions, although indirect interactions can also be detected (Barrios-Rodiles *et al*, 2005; Blasche & Koegl, 2013; Taipale *et al*, 2014).

Limitations A major disadvantage of the LUMIER method is that it requires lysis of cells prior to immunoprecipitation, a process that can result in the disruption of weak and transient PPIs, as well as the introduction of potential artifacts (e.g., by bringing together proteins in the lysate, which might not normally interact with one another in the cell, destabilizing proteins and exposing previously concealed non-native binding surfaces). The LUMIER assay must be carefully controlled, to normalize for differences in transfection efficiency and expression, and minimize background signal. The assay is not ideal for studying how PPIs change spatially, over time or in response to different environmental conditions (Barrios-Rodiles *et al*, 2005; Blasche & Koegl, 2013).

Mammalian protein–protein interaction trap (MAPPIT)

Principle The MAPPIT assay (Fig 1D) is designed for use in mammalian cell lines and is based on a cytokine signal transduction mechanism. A "bait" protein is fused to the C-terminus of a cytokine receptor deficient in binding to STAT3 (required for signal transduction), while "prey" proteins are fused to receptor fragments containing functional STAT3 recruitment sites. An interaction between a bait and prey proteins produces a functionally competent receptor, which, in response to cytokine ligand stimulation, activates STAT3 molecules (through intermediate JAK kinase activity), allowing them to enter the nucleus and induce transcription of a reporter system (e.g., luciferase; Ulrichts *et al*, 2009).

Advantages Mammalian protein–protein interaction trap provides a powerful way to examine mammalian PPIs directly in the context of the mammalian cell and is suitable for use in both HT library and array screening formats. The assay is easy to perform and does not require specialized equipment, beyond the necessary cell culture reagents and instrumentation to measure bioluminescence or fluorescence. Note that this method is best suited for studying binary interactions (Lievens *et al*, 2009, 2011). Variations of MAPPIT are effective for use in small-molecule screening approaches (Eyckerman *et al*, 2005; Caligiuri *et al*, 2006; Lievens *et al*, 2011).

Limitations Anchoring of the interaction sensor (i.e., the cytokine receptor) to the plasma membrane requires that PPIs occur in the

 © 2015 The Authors

cytoplasmic submembrane region, preventing detection of interaction with preys localized to other subcellular compartments. This anchoring (and the large size of the bait tag) may also block certain PPIs due to steric issues beyond those occurring in many other methodologies (Lievens *et al*, 2009). Finally, the method is also not compatible with full-length transmembrane proteins and is not suitable for spatial or temporal analysis of PPIs.

Kinase substrate sensor (KISS)

Principle Kinase substrate sensor (Fig 1E) is a recently developed mammalian two-hybrid approach designed to measure intracellular PPIs. In this assay, a "bait" protein is fused to the kinase domain of TYK2, while "preys" are coupled to a gp130 cytokine receptor fragment carrying TYK2 substrate motifs. Interaction of bait and prey results in phosphorylation of gp130 by TYK2, resulting in docking and activation of STAT3, which can then enter the nucleus and activate transcription of a STAT3-dependent reporter system (e.g., luciferase; Lievens *et al*, 2014).

Advantages Kinase substrate sensor allows assessment of PPIs directly in living mammalian cells and is sensitive enough to detect dynamic changes in response to physiological or pharmacological challenges. The method is effective for use with both membrane and cytosolic proteins and is best suited for measuring binary interactions (Lievens *et al*, 2014).

Limitations Like many other assays, the KISS readout is indirect, preventing spatial or temporal analysis of PPIs. The assay relies on endogenous STAT3, making this approach unsuitable for studying interactions involving proteins or stimuli that affect STAT3 signaling (Lievens *et al*, 2014).

Bimolecular fluorescence complementation (BiFC)

Principles Bimolecular fluorescence complementation (Fig 1F) is based on the division of a fluorescent protein (e.g., YFP) into two distinct non-fluorescent fragments, which are then fused to "bait" and "prey" proteins of interest. Interaction between bait and prey allows the two non-fluorescent fragments to associate and form a fluorescent complex, which can be viewed by microscopy or flow cytometry (Kerppola, 2008; Zhang *et al*, 2015).

Advantages Bimolecular fluorescence complementation allows direct visualization of PPIs in living cells, providing spatial information about the subcellular location where PPIs are occurring. The method is highly sensitive and can be used to detect interactions between proteins expressed at endogenous or near-endogenous levels, as well as weak and transient interactions. The method can be used for different organisms, is simple to set up, and is cost-effective. Different fluorescent proteins can also be used in combination, allowing the visualization of multiple PPIs in parallel in single cells. The method is best suited for detecting binary interactions (Hu *et al*, 2002; Kerppola, 2008; Zhang *et al*, 2015).

Limitations Bimolecular fluorescence complementation is not ideal for measuring PPI dynamics or real-time changes, due to a delay in generation of fluorescence upon protein interaction, as well as the irreversible nature of fluorochrome formation (Kerppola, 2008). Another disadvantage of BiFC includes functionality of fusion proteins, as is the case for other techniques involving protein tagging. Lastly, in some cases false-positive fluorescent signals can be detected by BiFC due to fluorescence intensity of reconstituted fragments arising irrespective of (or from non-specific) interaction between two proteins under investigation (Miller *et al*, 2015).

Mammalian membrane two hybrid (MaMTH)

Principle Mammalian membrane two hybrid (Fig 1B) is a recently developed *in vivo* proteomics technology designed for the analysis of mammalian membrane PPIs. The assay is based on the principle of split-ubiquitin, wherein reconstitution of inactive fragments of ubiquitin (Nub and Cub) upon interaction of proteins to which they are fused leads to release of an artificial transcription factor, and subsequent expression of a reporter system (luciferase in the case of MaMTH; Petschnigg *et al*, 2014).

Advantages Mammalian membrane two hybrid allows the analysis of the interactions of full-length mammalian membrane proteins directly in their natural cellular context. The assay is low cost, highly scalable, and readily transferable to virtually any cell line of interest. No specialized equipment is required, beyond standard cell culture reagents and tools necessary for monitoring luciferase activity. One of the key advantages of MaMTH is its high sensitivity, making it suitable for both the measurement of weak/transient interactions, and for monitoring dynamic, "condition-dependent" PPIs (i.e., which change in response to agonist, phosphorylation state, mutation etc.). The method is best suited for the detection of binary PPIs (Petschnigg *et al*, 2014).

Limitations For MaMTH to function, the bait must be associated with the membrane or other intracellular structures, to prevent non-specific activation of the reporter system (note that like MYTH, preys can be either soluble or membrane-bound). Additionally, the termini of the membrane protein fused to Cub must be cytosolic, in order to provide access to the deubiquitinating proteases responsible for cleavage and release of transcription factor. The method is also not suitable for spatial or real-time temporal analysis of PPIs (Petschnigg *et al*, 2014).

Fluorescence resonance energy transfer (FRET)

Principle Fluorescence resonance energy transfer (Fig 1G) is based on the non-radiative transfer of energy from an excited donor fluorophore to a nearby acceptor molecule. Donor and acceptors are selected such that the absorption spectrum of the acceptor fluorophore overlaps with the emission spectrum of the donor. In this approach, one protein of interest is fused to the donor, while the other is fused to the acceptor. If the two proteins interact or come into close proximity with one other, the donor and acceptor fluorophores are also brought together. Excitation of the donor in this case does not lead to photon release, but rather energy transfer to the nearby acceptor, which in turn produces an emission signal. This emission signal is distinct from the signal that would be observed for donor alone, and is used to monitor PPI (Ma *et al*, 2014).

Advantages A major advantage of FRET is its ability to monitor instantaneous, real-time PPIs, allowing the measurement of short-lived transient interactions. In addition, FRET can be used directly in the context of live cells and allows detection of interaction sites. Also,

© 2015 The Authors

due to the reversible nature of the fluorophore interaction, complex interaction dynamics can be monitored such as the dynamic equilibrium between complex formation and dissociation (Ma *et al*, 2014).

Limitations For FRET to function, protein fusions to appropriate fluorophores need to be generated (the technical demands of which may vary depending upon the fluorophores selected). In addition, for a strong FRET readout, close spatial proximity of the fluorophores is required for the energy transfer to occur. FRET also has decreased sensitivity compared to other fluorescence-based approaches like BiFC or BRET, as there tends to be strong background autofluorescence in cells upon sample illumination. For this reason, many controls are necessary to quantify the changes in fluorescence intensity in the presence and absence of energy transfer, and particularly weak interactions producing a signal close to background may be difficult to detect. Depending upon the fluorophores selected, photobleaching can also result in loss of signal over time (Boute *et al*, 2002; Ma *et al*, 2014).

Bioluminescence resonance energy transfer (BRET)

Principle The BRET assay (Fig 1G) has been developed to diminish a major limitation of FRET—the strong background signal that results from the direct excitation between the donor and acceptor fluorophores. In BRET, a protein of interest is fused to *Renilla* luciferase ("RLuc", serving as the energy donor), while its interacting partner is fused to either green or yellow fluorescent protein (GFP or YFP, serving as the energy acceptor). When donors and acceptors are brought into close proximity (< 100 Å) by interaction of their fusion partners, energy transfer occurs, producing fluorescent signal which is monitored to detect the PPIs (Boute *et al*, 2002; Hamdan *et al*, 2006).

Advantages Like FRET, BRET is able to monitor instantaneous real-time PPIs, functions directly in the context of live cells, and provides information about the cellular location at which an interaction occurs (Boute *et al*, 2002; Hamdan *et al*, 2006; Xie *et al*, 2011). However, BRET also has greater sensitivity than FRET, with lower background (Boute *et al*, 2002).

Limitations The major limitations of BRET are similar to those of FRET, including the need for the generation of fusion proteins, and the efficiency of the assay being dependent on close spatial proximity of the donor and acceptor (in order for proper energy transfer to occur; Hamdan *et al*, 2006). BRET signal also tends to be significantly weaker than that produced by FRET (Hamdan *et al*, 2006; Xie *et al*, 2011). In addition, the analysis of PPIs using BRET and FRET is not as easily scalable to HT screening applications as other methods, making it better suited to screens involving a more limited number of potential hits.

Affinity purification–mass spectrometry (AP-MS)

Principle Affinity purification–mass spectrometry (Fig 1H) is a popular technology that has gained considerable attention over the past decade. The general principle involves immobilization of "bait" protein of interest on a solid support (most frequently agarose or magnetic beads), and use of this coupled "bait" to capture target protein(s) from a soluble phase. Once affinity-purified, captured proteins are usually digested with proteases (e.g., trypsin), to generate peptides, which in turn are sub-fractionated using high-pressure liquid chromatography (HPLC) and then ionized and detected using a mass spectrometer. AP-MS can be conducted either with endogenous, native protein baits (using specific antibodies raised against them) or with protein baits to which a standardized "epitope tag" (e.g., TAP-, FLAG-, c-myc-, HA-, His-, protein A-, Strep-Tag) is fused. The choice of the most appropriate affinity purification method depends on a combination of factors, including the availability of antibodies, the type of a protein under investigation, and the scale of the conducted analysis (Dunham *et al*, 2012).

Advantages Affinity purification–mass spectrometry is a library-independent method with true genomewide HT capability. The main advantage of affinity purification using native antibodies against endogenous, native "baits" is that the proteins are purified in their natural form from cell or tissue lysates, eliminating issues associated with protein tagging and allowing multiple isoforms to be interrogated simultaneously. Conversely, the main advantages of epitope tagging are that it allows the study of proteins for which native antibodies are not available, and the analysis of multiple proteins using a single, defined process with a specific antibody (i.e., since many different proteins can be tagged with a single epitope; Dunham *et al*, 2012).

Limitations The major limitation of any AP-MS approach is a need to perform cell lysis and affinity purification. These steps do not allow for the detection of spatial or temporal PPIs and can also prevent detection of weak, transient PPIs. Another major limitation is contamination by abundant proteins co-purified from AP (Dunham *et al*, 2012) and artifacts resulting from exposure of proteins to one another in the unnatural environment of a cellular lysate (e.g., spurious interactions, disruption of protein interactions). In cases where epitope tags and ectopic expression are required, high background can also result from improper folding and mislocalization. However, several strategies do exist, which may help overcome these problems, such as use of appropriate negative controls, further enrichment of true interactions using tandem affinity purification (TAP), quantification approaches (SILAC and other isotopic labeling, and label-free quantification, which usually need special computational tools; Choi *et al*, 2011) and, for smaller experiments, contaminants can be filtered out by comparison with a contaminant repository database (Mellacheruvu *et al*, 2013). With endogenous proteins, low expression levels of proteins of interest may also prevent detection (Dunham *et al*, 2012). Lastly, data analysis of AP-MS experiments is more difficult compared to other PPI assays (i.e., Y2H), due to required expertise with MS and specific bioinformatics tools needed to address the limitations listed above.

Proximity-dependent biotin identification coupled to mass spectrometry (BioID-MS)

Principle Proximity-dependent biotin identification coupled to mass spectrometry (Fig 1I), which is similar in nature to AP-MS, uses a "bait" protein of interest fused to a prokaryotic biotin ligase molecule (BirA). When expressed in cells, proteins in proximity to the BioID fusion protein are biotinylated by BirA, permitting their selective isolation using an avidin/streptavidin-based biotin affinity capture approach. These purified, biotinylated proteins are then identified using MS, providing a list of candidate interacting partners for the bait of interest (Roux *et al*, 2012).

 © 2015 The Authors

Advantages Similar to AP-MS, BioID-MS is also a library-independent method. A major advantage of BioID-MS is that PPIs are detected in their natural cellular context (since biotinylation occurs in the cell prior to lysis). Additionally, issues associated with bait/prey stability and disruption of interactions upon cell lysis are avoided. The method is also well suited for identifying weak or transient interactions and is amenable to temporal regulation (potentially allowing for pulse-chase type applications; Roux *et al*, 2012). The method also appears to be more effective at detecting low-abundance proteins than AP-MS (Lambert *et al*, 2015).

Limitations BioID-MS requires the fusion of bait protein to BirA, which adds significantly to the size of the protein and can potentially compromise its targeting or function. Low expression level of PPI partners can also lead to false negatives. The biotinylation process itself may also affect protein behavior/interactions in certain cases (Roux *et al*, 2012). Finally, like AP-MS, required expertise in MS and specific bioinformatics tools can make data analysis more complex than for other methods.

Proximity ligation assay (PLA)

Principle Proximity ligation assay (Fig 1J) is a powerful method for the *in situ* detection of PPIs in fixed cells and tissues. The general premise of the assay involves the use of proximity probes (i.e., antibodies conjugated with DNA oligonucleotides, which are able to recognize two target proteins of interest). When the Proximity Probes are brought close to one another (i.e., due to interaction of the target proteins to which they are bound), the DNA strands serve as a template to direct ligation of two subsequently added oligonucleotide fragments into a circular molecule. This circular DNA is then amplified using rolling circle amplification (RCA) primed by one of the original proximity probe oligonucleotides, resulting in a long DNA sequence physically linked to the corresponding antibody (and thus the interacting protein pair). This new DNA sequence contains many repetitive elements, which are then bound by fluorophore-labeled complementary oligonucleotide probes, allowing visualization of interactions, at the specific sites where they occur, using a fluorescence microscope (Koos *et al*, 2014).

Advantages The major advantages of PLA are its ability to detect and localize PPIs with single molecule resolution and objectively quantify them in cells and tissues. In addition, transient or weak interactions can be monitored (Koos *et al*, 2014).

Limitations A major disadvantage of PLA is the dependence on enzymes (i.e., for ligation and polymerization), making the approach expensive and highly dependent on enzyme activity and stability. The required use of antibodies in PLA is another potential drawback, as antibodies are often costly, and may also not be readily available against particular proteins of interest (Koos *et al*, 2014). Thus, PLA is not ideally suited for HT PPI screening applications.

Ligand–receptor capture – trifunctional chemoproteomics reagents (LRC-TriCEPS)

Principle The LRC-TriCEPS approach (Fig 1K) has been developed to elucidate potential receptor/ligand interactions. TriCEPS employs a chemoproteomics reagent consisting of three moieties—one that binds ligands of interest containing an amino group, a second that binds glycosylated receptors on live cells, and a biotin tag. The reagent effectively serves as a stable "bridge", covalently linking a ligand of interest to carbohydrate groups on its cognate receptor. Following treatment with TriCEPS, cells are lysed and enzymatically digested with trypsin, and TriCEPS bound peptides are purified via the biotin tag. Receptor peptides are then freed from the TriCEPS reagent and identified using quantitative MS (Frei *et al*, 2012, 2013).

Advantages The major advantages of using LRC-TriCEPS are the ability to detect ligand–receptor interactions without the need for genetic manipulations. This approach is also effective for detecting surface interactions that are very transient in nature, and can be used with both populations of individual cells and tissue samples. In addition, LRC-TriCEPS can be used to identify the cell surface binding partners of many different types of ligands, including peptides, proteins, viral particles, antibodies, and engineered affinity binders (Frei *et al*, 2012, 2013).

Limitations By design, LRC-TriCEPS is only useful for identifying N-glycoprotein receptors, and is ineffective if glycans are sterically inaccessible. Coupling of ligands to TriCEPS reagent may also affect their functionality/proper target binding in some cases (necessitating verification of ligand function following TriCEPS linkage, where possible). TriCEPS may also not be effective in detecting receptor–ligand interactions in situations where ligand binding requires association to other cell surface structures (in addition to a target glycoprotein; Frei *et al*, 2013).

Avidity-based extracellular interaction screen (AVEXIS)

Principle Avidity-based extracellular interaction screen is a PPI assay developed to systematically screen for novel extracellular receptor–ligand pairs involved in cellular recognition processes (Fig 1L). The general premise of this approach involves the expression of secreted recombinant "bait" and "prey" proteins (e.g., natively secreted proteins or the truncated ectodomain of membrane proteins containing an N-terminal secretory peptide) in a mammalian cell-based system so that structurally important posttranslational modifications can occur. Bait proteins are biotinylated, so they can be captured on a streptavidin-coated solid phase, while prey proteins are tagged with β-lactamase and a peptide sequence directing their pentamerization (used to increase effective prey concentration and improve assay sensitivity). Bait and prey isolates are then presented to one another in a binary manner to detect direct interactions using an ELISA-type format (Kerr & Wright, 2012).

Advantages Avidity-based extracellular interaction screen can detect very weak PPIs, which are a typical feature of interactions between membrane-embedded receptor proteins (it has been shown to detect interactions with equilibrium dissociation constants as low as ~10 μM) with a low false-positive rate (Sun *et al*, 2012; Kerr & Wright, 2012). The assay has also been adapted for use on a higher-throughput scale than many other assays designed to detect extracellular interactions (Sun *et al*, 2012).

Limitations Avidity-based extracellular interaction screen is limited to the study of membrane proteins with self-contained

© 2015 The Authors

extracellular domains and is not generally suitable for multipass membrane proteins and other proteins that need to be embedded in the plasma membrane to fold and function properly (Kerr & Wright, 2012; Frei *et al*, 2013). In addition, selecting, preparing, and validating the constructs necessary for AVEXIS can be a lengthy (and relatively costly) process, although the use of a recently reported protein microarray format does help in this regard (Sun *et al*, 2012). The approach may also have difficulty detecting homophilic interactions and is not ideal for quantitatively comparing the strength of different interactions (due to the artificial pentamerization of preys; Kerr & Wright, 2012).

Dynamic protein interaction networks

A major limitation of the available PPI interaction network data is the static representation of these interactions, neglecting the temporal and spatial organization of protein dynamics as well as the effect of posttranslational modifications (PTMs). For instance, a PPI may occur only during specific time periods (e.g., under particular stress conditions, in response to certain signaling events etc.) or if specific PTMs are present. The nature of protein–protein interaction is thus an inherently dynamic process that changes with time, environments, and at different stages of the cell cycle. Recently, dynamic protein interaction networks have been constructed by using proteomic, genomic, and transcriptomic methodologies. In the previous section, we touched briefly on the suitability of some techniques for mapping dynamic interactions. Examples of some of specific proteomic-based approaches employed include Y2H and AP-mass spectrometry (Woodsmith & Stelzl, 2014). For instance, the phosphotyrosine-dependent PPI network was recently studied using Y2H, and identified many novel phosphotyrosine-dependent PPIs of human kinases (Grossmann *et al*, 2015). AP-MS approaches have been employed to study the dynamics of the human 26S proteasome-interacting proteins (Wang & Huang, 2008, 2014), study changes in the interactome of 14-3-3β in response to activation of the insulin-PI3K-AKT pathway (Collins *et al*, 2013), and map phosphotyrosine-dependent interaction sites on ErbB-receptor family members (Schulze *et al*, 2005).

In addition to the temporal and spatial organization of PPIs, perturbations of PPIs from disease-associated alleles have also gained much interest. Such perturbations can be either subtle or dramatic, but often have significant biological consequences, and understanding the nature of these changes can be important in developing new therapeutic strategies. Thousands of genetic variants have been identified in many Mendelian disorders, complex traits, and cancers; however, the effect of these genetic variants on PPI networks is still far from clear. Recent studies have looked into assessing perturbations of protein interactions by disease-associated alleles using the techniques described above. For example, Sahni *et al* (2015) found widespread perturbations of macromolecular interactions caused by disease-specific mutant alleles using a comprehensive genomics/proteomics approach involving LUMIER and Y1H/Y2H technologies. Additionally, Wang *et al* (2012) integrated available protein structure and large-scale PPI data to comprehensively investigate the relationships between mutations, protein interactions, and human disease. Work by Petschnigg *et al* (2014) using the MaMTH assay also identified differential interactors of WT and oncogenic mutant forms of the receptor tyrosine kinase EGFR. For a more thorough examination of protein interactome networks and disease, several excellent reviews are available (Ideker & Sharan, 2008; Vidal *et al*, 2011; Sahni *et al*, 2013).

Analysis of PPI screen data

Once a screen is completed, it is necessary to properly analyze the data in order to validate the results and improve overall interactome quality. In this section, we provide an overview of some key considerations and methods useful for analysis of PPI datasets.

Assessing PPI datasets

Assessing the frequency of false positives and false negatives in PPI datasets has been a long-standing problem, especially for HT screens. Typically, the frequency of false positives is measured as the false discovery rate ($FDR = \frac{false\ positives}{true\ positives + false\ positives}$), and the frequency of false negatives as the false negative rate ($FNR = \frac{false\ negatives}{true\ positives + false\ negatives}$) or sensitivity ($\frac{true\ positives}{true\ positives + false\ negatives}$). The main strategies for assessing FDR and sensitivity have involved testing detected interactions by multiple methods and comparing against interactions from literature. FDR has been, arguably, a greater focus in interactome studies than sensitivity. In small-scale screens, FDR can be assessed and minimized by testing all reported interactions using multiple methods. However, the FDR of small-scale screens may still be uncertain. Edwards *et al* (2002) found that interactions from small-scale screens were not always consistent with known 3D structures of protein complexes. Rolland *et al* (2014) tested interactions reported in single publications and found that the detection rate was just slightly higher than for random protein pairs. Small-scale studies rarely report protein pairs that were tested but not detected. Consequently, the sensitivity of their screens and the assessment of how much of the interactome they have tested are largely unknown (Cusick *et al*, 2009).

Assessing HT screens is more difficult since testing all detected interactions by multiple methods is not feasible. Venkatesan *et al* (2009) developed a rigorous framework for assessing the quality of HT PPI datasets. Their framework calculates four parameters: screening completeness, assay sensitivity, sampling sensitivity, and precision. Screening completeness is the fraction of open reading frame (ORF) pairs tested in the screen. Assay sensitivity is the fraction of interactions that can be identified by the assay, estimated by testing the assay on a gold standard set of interactions, and determining the fraction detected. Sampling sensitivity is the fraction of detectable interactions identified in one trial of the assay, estimated by repeating the assay multiple times and fitting a Bayesian model to the results. Precision, the fraction of detected pairs that are true positives, can be estimated by testing the assay on reference sets of interacting and non-interacting protein pairs, and calculating the fraction of detected pairs that are from the interacting set.

Once the precision and sensitivity of an assay have been estimated, the assay can be used to determine the FDR of interaction datasets. HT studies commonly estimate FDR by retesting a subset of detected protein pairs using different small-scale or HT methods (Yu *et al*, 2008; Simonis *et al*, 2009; Rolland *et al*, 2014). Such estimates need to take into account the precision and sensitivity of the

 © 2015 The Authors

retesting assay. This is especially important as the precision or sensitivity may be quite limited. Braun *et al* (2009) assessed five HT assays on a gold standard dataset comprised of interactions reported by multiple small-scale studies (positive cases) and an equal number of randomly chosen protein pairs. Each HT assay detected only ~20–35% of positive cases, and up to 4% of negative cases. Combined, the five assays detected 59% of positive cases, while FDR increased to 14%. Furthermore, different assays have different systematic biases; for example, affinity purification methods may be biased in favor of high-abundance proteins (Ivanic *et al*, 2009). Another concern, especially difficult to address, is that gold standard datasets may also have biases. Such datasets, often comprising PPIs reported by multiple small-scale studies, may be deficient for certain types of proteins or PPIs due to research bias or limitations of assays (Hakes *et al*, 2008; Edwards *et al*, 2011; Rolland *et al*, 2014; Kotlyar *et al*, 2015; Wang *et al*, 2015).

Computational methods for assessing results

Computational methods provide a means of estimating FDR without retesting detected protein pairs. Furthermore, they can provide error estimates for specific proteins or protein pairs.

D'haeseleer and Church (2004) introduced a method for estimating FDR and assessing the reliability of individual interactions. Their method analyzes the overlap of detected interactions with two other datasets, including a trusted reference set. The variance of FDR estimates can be high if the overlaps are small. However, the authors showed that FDR estimates are not greatly affected by the quality of the chosen datasets.

A statistical model introduced by Huang and Bader for assessing two-hybrid datasets provides global error estimates as well as error estimates for specific baits (Huang & Bader, 2009). Thus, it can determine whether certain baits are responsible for a disproportionate share of the global error rate. For example, in two-hybrid data from worm, it found that the FDR was especially high among proteins involved in cellular metabolic processes.

Computational methods to help improve data quality

One approach for assessing and improving data quality is to examine whether a PPI dataset possesses properties of interacting protein pairs. Methods that use this approach assume that interacting proteins are likely to have co-expressed genes (Deane *et al*, 2002), shared subcellular localization (Sprinzak *et al*, 2003), similar functional and process annotations (Sprinzak *et al*, 2003; Wang *et al*, 2007), and shared interaction partners (Saito *et al*, 2002; Goldberg & Roth, 2003). Evaluating a PPI dataset using such evidence can be problematic: Many interacting protein pairs do not have correlated gene expression, protein annotations such as subcellular localization and function are often incomplete or unavailable, and shared interaction partners are frequently unknown, since the interactomes of most species are largely unmapped. However, ranking detected protein pairs using these types of evidence can help identify true positive interactions. A combination of such evidence has been used in HT studies to define high-confidence (HC) subsets of detected interactions (Miller *et al*, 2005; Havugimana *et al*, 2012). The evidence can also help identify potential false negatives—protein pairs that are not strongly supported by the experimental detection method but have properties of true interactions. Unfortunately, ranking based on this evidence can introduce biases; ranking by correlation of gene expression profiles favors stable interactions (Brown & Jurisica, 2007), while ranking by shared Gene Ontology terms or interaction partners favors well-studied proteins. If the evidence is used multiple times during the planning and analysis of an experiment, there may be a danger of circular reasoning. For example, testing protein pairs with similar functions and then ranking detected pairs by similarity of localizations would be largely ineffective, as functional similarity is correlated with localization similarity.

An approach for improving the quality of AP-MS data involves calculating a score for each co-purified pair, indicating the likelihood of the two proteins being observed together. Such scores have been calculated using various methods: log-ratios of observed versus expected co-occurrences (Gavin *et al*, 2006), machine-learning algorithms (Krogan *et al*, 2006; Collins *et al*, 2007), hypergeometric probabilities (Hart *et al*, 2007), and randomizations (Yu *et al*, 2009). Scores can be used for ranking protein pairs and defining a HC subset of interactions. A score threshold for defining this subset can be determined based on FDR and sensitivity calculated from a gold standard dataset comprising interacting and non-interacting protein pairs. True positive interactions can be distinguished from contaminants by analyzing quantitative information from mass spectrometry data, including spectral counts, signal intensity in the precursor scan of the mass spectrometer, and intensity of product ions after fragmentation (Gingras & Raught, 2012). Analysis of quantitative data using tools such as SAINT (Choi *et al*, 2011) can be especially helpful when aiming to detect transient interactions; preservation of transient interactions in AP-MS experiments requires short incubation times and few washes, resulting in more contaminants that need to be filtered (Gingras & Raught, 2012).

Varjosalo *et al* (2013) filtered AP-MS data in three steps to remove different types of protein contamination. The first filter removed proteins that may have been left over from a proceeding experiment. The second filter removed non-specifically interacting proteins. The third filter removed low-abundance, non-systematic contaminants; bait–prey pairs were assigned weighted spectral-count-based scores reflecting interaction abundance and reproducibility, and pairs with scores below a threshold were removed. This three-step filtering improved reproducibility of resulting networks more than previous filtering methods, wD-score (Behrends *et al*, 2010) and SAINT (Choi *et al*, 2011).

Computational prediction to help improve datasets and the interactome

Computational PPI prediction is similar to previously described methods that assign scores to detected protein pairs, indicating the likelihood of interaction. However, prediction methods provide scores for both detected and undetected PPIs. These scores can be used to improve the quality of experimentally detected PPIs, by identifying high-confidence subsets and potential false negatives, which can help accelerate interactome mapping (Schwartz *et al*, 2009). Also, prediction methods can help fill the gaps in a known interactome by predicting interactions for interactome orphans and low-degree proteins (Kotlyar *et al*, 2015).

PPI prediction methods can be categorized by the data they use: genomic data, protein sequence, protein structure, PPI networks, gene expression, and annotations of gene function,

© 2015 The Authors

localization, and process. Prediction methods based on genomic data analyze conserved operon structure, fusion domains, phylogenetic profiles, and interologs. Analysis of operon structure is based on the idea that genes in close proximity on the genome are more likely to encode interacting proteins, especially if the proximal locations are conserved across species (Dandekar *et al*, 1998; Overbeek *et al*, 1999). Similarly, if two genes exist as a single fused gene in another species, they are likely to encode interacting proteins (Enright *et al*, 1999). Phylogenetic profiles are used to identify gene pairs that tend to co-occur across species—either both are present in a species, or both are absent; such pairs are likely to be functionally related and may encode interacting proteins (Pellegrini *et al*, 1999). Interologs are interactions conserved across species: if a pair of proteins interacts in one species, their orthologs in another species are more likely to interact (Walhout *et al*, 2000; Yu *et al*, 2004).

Many studies have predicted interactions based on protein sequence (Gomez *et al*, 2003; Martin *et al*, 2005; Nanni & Lumini, 2006; Shen *et al*, 2007; Guo *et al*, 2008; Zaki *et al*, 2009; Chang *et al*, 2010; Guo *et al*, 2010; Yu *et al*, 2010). These studies analyze experimentally determined interactions to find patterns that distinguish sequences of an interacting protein pair from those of a non-interacting pair. This is often done using machine-learning algorithms such as support vector machines, random forests (Roy *et al*, 2009), and K-local hyperplane nearest-neighbors (Nanni & Lumini, 2006). Protein sequence can also be used indirectly for prediction; protein domains can be determined from sequence, and pairs of domains enriched among known interacting protein pairs may predict new interactions (Sprinzak & Margalit, 2001; Wojcik & Schächter, 2001; Nguyen & Ho, 2006; Singhal & Resat, 2007). Another approach combines sequence with protein tertiary structure. Although few proteins or complexes have known 3D structure, sequence homology can serve as a link to other proteins. Homologous proteins, especially with conserved binding sites, are likely to interact in similar ways (Aloy & Russell, 2003; Ma *et al*, 2003; Sinha *et al*, 2010). Thus, a protein pair can be predicted to interact based on sequence or structural homology to proteins in solved complexes (Lu *et al*, 2002; Aloy & Russell, 2003; Zhang *et al*, 2012a).

If an interactome is partially known, new interactions may be predicted from known network structure, often based on the idea that interacting proteins tend to share interaction partners (Saito *et al*, 2002; Goldberg & Roth, 2003; Liu *et al*, 2008). Other types of interaction evidence—including correlated gene expression, shared subcellular localization, similar function and process—are typically used in combination with other evidence (Jansen *et al*, 2003; Ben-Hur & Noble, 2005; Rhodes *et al*, 2005; Elefsinioti *et al*, 2011).

PPI prediction methods have a number of limitations and biases, often similar to those of experimental PPI assays. Computational methods tend to have difficulty predicting transient interactions and interactions involving lesser studied proteins, which typically have no tertiary structure data, no detailed Gene Ontology or domain annotations, few known interactions, and few orthologs in different species (Kotlyar *et al*, 2015). Transient interactions are difficult to predict based on correlation of gene expression profiles (Brown & Jurisica, 2007), or analysis of protein sequence or structure. In these interactions, the two encoding genes are not highly correlated (Brown & Jurisica, 2007), interaction interface sequences are not as conserved as in obligate interactions (Perkins *et al*, 2010), interacting proteins often undergo conformational changes (Perkins *et al*, 2010), and interactions are frequently mediated by linear motifs rather than globular domains (Perkins *et al*, 2010). If proteins lack Gene Ontology annotations, known interaction partners, or orthologs, it is difficult to predict their interactions using annotation similarity, network topology, or comparative genomics, respectively. Interestingly, such proteins are also underrepresented in the experimentally detected human interactome (Kotlyar *et al*, 2015). If prediction methods require training, they may acquire the biases of their experimentally detected training set. This may explain the finding of Rolland *et al* that a proteome-wide, structure-focused prediction method, PrePPI (Zhang *et al*, 2012a), had a tendency to report interactions among well-studied proteins (Rolland *et al*, 2014).

Databases: what is available and what do they tell us?

Studies that detect PPIs report their findings in journal articles as free-form text. Consequently, original information about detected PPIs is scattered across thousands of articles and requires manual curation. Converting these data into an easily usable set of interactions and experimental descriptions remains a daunting problem, comprising several key tasks: (i) experiments need to be described with a controlled vocabulary and recorded in a common format, (ii) thousands of articles have to be curated and resulting information has to be easily accessible, and (iii) proteins need to be unambiguously identified. The first task, creation of standard vocabularies and formats for PPI data, was addressed through the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) by the Molecular Interaction (MI) workgroup. They created a common controlled vocabulary for experimental techniques, molecular features, and interaction types (Orchard & Kerrien, 2010), and XML (PSI-MI XML) and tab-delimited (MITAB) formats for recording and transferring data (Kerrien *et al*, 2007). Most major PPI databases adopted the vocabulary and data formats, allowing users to easily integrate datasets and analyze them with programs such as Cytoscape (Su *et al*, 2014) and NAViGaTOR (Brown *et al*, 2009).

The second task, curating articles and providing results through online databases, started with the DIP (Salwinski *et al*, 2004) and BIND (Bader *et al*, 2003) database projects and has continued with the creation of many similar resources (Table 2 and Fig 2A). These resources were especially important given the rapid increase in the number of human PPIs (Fig 2B) detected by various experimental methods. Initially, the focus was on experimentally detected PPIs in yeast and human, but the scope has greatly expanded. Some databases now include computationally predicted interactions (e.g., STRING (Szklarczyk *et al*, 2015), FpClass (Kotlyar *et al*, 2015), IID (Kotlyar *et al*, 2016)), functionally related protein pairs (e.g., STRING (Szklarczyk *et al*, 2015)), interactions between proteins and other molecule types (e.g., BIND (Bader *et al*, 2003), BindingDB (Liu *et al*, 2007), IntAct (Kerrien *et al*, 2012)), interactions in a range of organisms (e.g., DIP (Salwinski *et al*, 2004), IntAct (Kerrien *et al*, 2012), MINT (Licata *et al*, 2012), BioGRID (Chatr-Aryamontri *et al*, 2015), IID (Kotlyar *et al*, 2016)), and interactions involving

 © 2015 The Authors

specific types of proteins (e.g., extracellular matrix—MatrixDB (Chautard *et al*, 2011), immune related—InnateDB (Breuer *et al*, 2013)).

Databases also differ in several other respects: the level of detail recorded about experiments (shallow versus deep), the way information is acquired (manual curation of literature or automatic approaches such as text mining or PPI prediction), and the sources of information—peer-reviewed articles (primary sources) or other databases (secondary sources). Manual curation of articles is the most trusted method for acquiring data and is carried out by most databases. Several approaches have been used to assist with curation: curation guidelines have been established, an automated syntax checker was implemented to test for compliance with accepted formats, and guidelines were created for reporting PPIs in papers, so that curation of papers is easier and more accurate (Orchard *et al*, 2007). However, manually curating all previously published literature and keeping up with new publications is a huge task. The IMEx consortium (Orchard *et al*, 2012) was created to better organize the curation effort across major PPI databases. Members of the consortium avoid curating the same papers and follow the same curation rules.

Although current PPI databases provide easy access to their PPI data, obtaining the most complete up-to-date network can be challenging: the latest data has to be downloaded from multiple databases and merged. The PSICQUIC (Aranda *et al*, 2011) query interface simplifies these tasks. It enables multiple PPI databases to be searched with the same query, and a clustering algorithm provided with PSICQUIC helps merge results by grouping interaction evidence based on primary identifiers.

Unfortunately, the third task, unambiguously identifying proteins, has not been entirely resolved. Most PPI databases use UniProtKB (Magrane & Consortium, 2011) protein identifiers, which can represent peptides, fusion proteins, specific isoforms, or proteins whose isoforms are not specified. However, some databases use Ensembl (Flicek *et al*, 2014), Entrez (Maglott *et al*, 2011), RefSeq (Pruitt *et al*, 2012), and species-specific identifiers. Mapping between different types of identifiers is not always possible. When databases use different protein identifiers, PSICQUIC is unable to identify redundant data. However, this problem is being addressed as more providers of PSICQUIC clients include commonly used identifiers in export files (Orchard, 2012).

Visualization, analysis, and biological validation of PPI data

Visualization Tools

A number of tools are available for visualization and analysis of PPI data including Cytoscape (Su *et al*, 2014), NAViGaTOR (Brown *et al*, 2009), and packages from R and Bioconductor (e.g., Rintact (Chiang *et al*, 2008) combined with RBGL). The main types of functionality supported by these tools include loading PSI-MI files, visualizing networks, annotating networks, and conducting network analysis. Data can be loaded from tab-delimited or PSI-MI XML files, and visualized as a graph, with nodes representing proteins and edges representing interactions. Multiple graph layouts are supported, including grids and force-directed layouts. Annotations for nodes and edges can be included in the original data files, retrieved from other text files, or imported from databases. The appearance of nodes and edges can be set based on annotations. Network analysis capabilities include clustering to identify protein complexes, motifs, or graphlets, calculating centrality measures, and identifying shortest paths or flows.

Identifying complexes in PPI networks

Although PPI networks focus on pairwise interactions, cellular processes are often carried out by protein complexes. Complexes typically have a "core"—a central functional unit present in most isoforms of the complex, and "attachments"—proteins present in some isoforms of the complex (Gavin *et al*, 2006). The attachments may include "modules"—sets of proteins that always appear together in different complexes (Gavin *et al*, 2006).

Experimental methods for detecting PPIs cannot easily identify complexes. For example, Y2H methods only identify binary interactions, and while TAP and HMS-PCI identify potential complexes, reliably identifying complex members requires "reverse purification"—repeatedly applying the detection method, using candidate members of the complex as baits (Gavin *et al*, 2002).

Table 2. Major protein–protein interactions (PPI) databases.

Database	Reference	URL	IMEx member	PPI evidence	Specialization
BioGRID	Chatr-Aryamontri *et al* (2015)	http://thebiogrid.org	Observer	Experimental	
DIP	Salwinski *et al* (2004)	http://dip.doe-mbi.ucla.edu/dip	Yes	Experimental	
FPCLASS	Kotlyar *et al* (2015)	http://ophid.utoronto.ca/fpclass	No	Computational	
HPRD	Keshava Prasad *et al* (2009)	http://www.hprd.org/	No	Experimental	
IID	Kotlyar *et al* (2016)	http://ophid.utoronto.ca/iid	Yes	Computational, Experimental	
InnateDB	Breuer *et al* (2013)	http://www.innatedb.ca	Yes	Experimental	Immune-related PPIs
IntAct	Kerrien *et al* (2012)	http://www.ebi.ac.uk/intact	Yes	Experimental	
iRefWeb	Turinsky *et al* (2014)	http://wodaklab.org/iRefWeb/	No	Experimental	
MatrixDB	Chautard *et al* (2011)	http://matrixdb.ibcp.fr/	Yes	Experimental	Extracellular matrix PPIs
MINT	Licata *et al* (2012)	http://mint.bio.uniroma2.it/mint	Yes	Experimental	
STRING	Szklarczyk *et al* (2015)	http://string-db.org	No	Computational, Experimental	Functional protein–protein associations

© 2015 The Authors

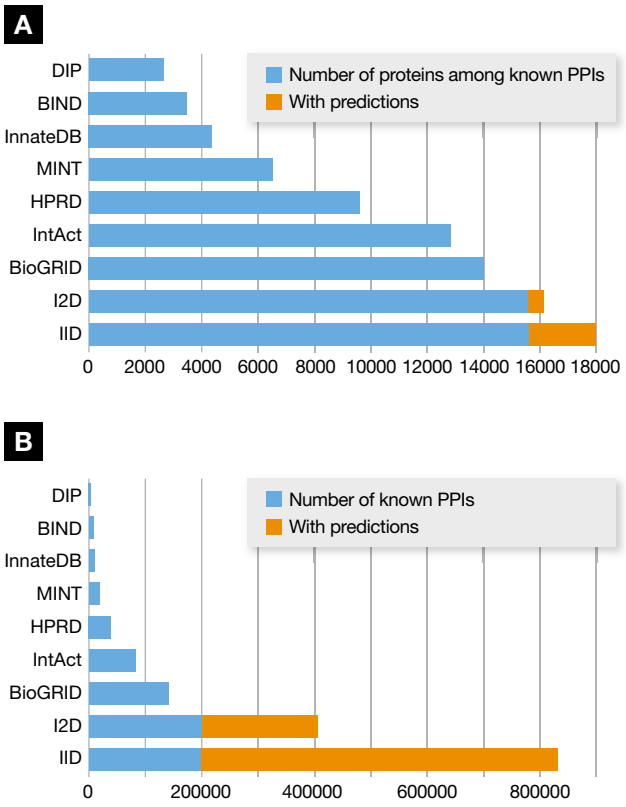


Figure 2. Protein and PPI counts in major human PPI databases.
(A) Major human PPI databases and the number of proteins they contain.
(B) Major human PPI databases and the number of PPIs they contain.

Computational methods can predict complexes by analyzing PPI networks and integrating networks with information such as gene function or co-expression. Most complex prediction methods share the same main steps: (i) assigning confidence scores to detected interactions, (ii) identifying complexes by clustering PPI networks or analyzing additional data, and (iii) evaluating resulting complexes by comparing with gold standard datasets. The first step assigns confidence scores to detected interactions; scores can be used to filter interactions or can be included as input to clustering algorithms. Any of the scoring approaches described earlier may be used (see section "Computational methods to help improve data quality"). Often, complexes are determined from AP-MS data, and scoring approaches specific to this data are used.

Most prediction methods assume that complexes correspond to highly connected regions of PPI networks and cluster the networks to identify these regions. The clustering approaches can be categorized as agglomerative or divisive, and overlapping or non-overlapping. Agglomerative approaches (Bader & Hogue, 2003; Li *et al*, 2005; Liu *et al*, 2009; Wang *et al*, 2009; Nepusz *et al*, 2012) start with seeds—individual nodes or cliques—and expand them into larger clusters by adding single nodes or merging with other clusters. Divisive approaches (van Dongen, 2000; Pu *et al*, 2007; Friedel *et al*, 2009) start with an entire network and partition it into highly connected regions. Overlapping approaches (Wang *et al*, 2009; Nepusz *et al*, 2012) allow nodes to be members of multiple clusters, to reflect overlap between complexes, while non-overlapping approaches (van Dongen, 2000; Bader & Hogue, 2003; Liu *et al*, 2009) assign nodes to single clusters. Some methods (Wu *et al*, 2009; Leung *et al*, 2009; Srihari *et al*, 2010; Chin *et al*, 2010) try to identify core and attachment sections of complexes.

Several methods combine network clustering with information about protein function, orthology, or structure. To increase the reliability of predicted complexes, these methods look for clusters whose members have similar functions (King *et al*, 2004; Li *et al*, 2007), highly conserved orthologs in the same set of species (i.e., the complex is conserved as a functional unit (Sharan *et al*, 2005; Hirsh & Sharan, 2007)), and protein structures enabling simultaneous interactions with multiple complex members (Ozawa *et al*, 2010; Jung *et al*, 2010). Several surveys (Brohée & van Helden, 2006; Vlasblom & Wodak, 2009; Li *et al*, 2010; Srihari & Leong, 2013) evaluated complex prediction methods by comparing their results against experimentally determined complexes. The Markov Cluster Algorithm (van Dongen, 2000; van Dongen & Abreu-Goodger, 2012) was found to be a top clustering method in three surveys (Brohée & van Helden, 2006; Vlasblom & Wodak, 2009; Li *et al*, 2010), and integration of network clustering with other information significantly improved performance (Srihari & Leong, 2013).

Identifying interaction conditions

Understanding how PPIs produce specific phenotypes requires information on their context: when, where, and under what conditions interactions occur. Computational methods can help determine this information by text mining of the PubMed database (Chowdhary *et al*, 2012), or more commonly, by integrating transcriptomic and other data with PPI networks. Usually, these methods aim to identify cell types, tissues, and disease states in which interactions occur.

Direct evidence for interactions occurring in a given cell type or tissue is often unavailable since PPI detection is typically done in yeast cells or common cell lines. By contrast, HT gene expression data are available for a wide variety of organisms, cell types, tissues, and conditions (Barrett *et al*, 2013; Kolesnikov *et al*, 2015). A common approach for assigning PPIs to tissues is to check whether the genes encoding an interacting protein pair are both expressed in a tissue (Bossi & Lehner, 2009; Lopes *et al*, 2011). Proteomics data (Uhlen *et al*, 2015) are less extensive but can be used analogously. The TissueNet (Barshir *et al*, 2013) database uses both gene and protein expression data to assign PPIs to tissues. Assigning tissues on the basis of gene or protein expression has limitations; absence of gene expression may not indicate absence of protein expression, and presence of gene or protein expression may not mean that proteins interact. Also, since this approach estimates the presence or absence of proteins, it can only indicate whether all interactions involving a protein are absent, but not whether a specific interaction is absent. Correlation of gene expression profiles can provide information on specific interactions; a pair of genes with correlated expression profiles in a tissue or cell type may have interacting protein products (Camargo & Azuaje, 2007).

Interactions that change in disease or other conditions can be identified by similar approaches. An interaction may be disease-related if the two encoding genes are both expressed only in the disease state (or are upregulated in the disease state; Ideker *et al*, 2002), or the genes have correlated expression profiles in disease states (Camargo & Azuaje, 2007; Guo *et al*, 2007; Xiao *et al*, 2012).

 © 2015 The Authors

Differential correlation of gene expression profiles can provide more specific information about an interaction: if two genes have significantly different correlation levels in two conditions, then the interaction of their protein products may change between conditions (Lin *et al*, 2010; Yoon *et al*, 2011; Zhang *et al*, 2012b; Yu *et al*, 2013).

Integrating PPI networks with other omics data

Integrating PPI networks with other omics data, such as genomic, transcriptomic, and proteomic, is essential for understanding the molecular basis of phenotypes. Just as gene expression data can provide context for PPIs, and link them with specific conditions, PPI networks can provide context for other data, and link it with phenotype.

One of the most common types of integration combines PPI networks and gene-phenotype data, to uncover relationships between genes and diseases. Goh *et al* (2007) showed that essential genes and disease genes have distinct network properties: essential genes tend to encode hub proteins, while disease genes encode proteins in the periphery of the network. Genes implicated in similar diseases tend to encode proteins that are close in PPI networks—either interacting directly (Goh *et al*, 2007; Schadt, 2009) or members of the same complex (Lage *et al*, 2007), pathway (Wood *et al*, 2007), or subnetwork (Lim *et al*, 2006). Based on this idea, it is possible to identify novel disease genes by mapping known disease-associated genes to nodes in PPI networks, and applying random walk (Kohler *et al*, 2008; Smedley *et al*, 2014), network flow (Yeger-Lotem *et al*, 2009; Chen *et al*, 2011), label propagation (Lee *et al*, 2011a), or other related algorithms (Vanunu *et al*, 2010; Winter *et al*, 2012). Random walk algorithms have been shown to be especially effective (Navlakha & Kingsford, 2010).

Integrating PPI networks with protein–DNA interactions, gene expression, phenotype, and drug information can provide insights into disease and drug mechanisms and can help identify new treatments. PPI networks combined with protein–DNA interactions have been used to model cellular regulatory networks—identifying regulatory circuits (Yeger-Lotem *et al*, 2004) and signaling-regulatory pathways (Ourfali *et al*, 2007). More recently, networks were used to predict disease mechanisms by modeling pathogen induced perturbations (Gulbahce *et al*, 2012), and effects of node or edge removal (Zhong *et al*, 2009; Sahni *et al*, 2013). Networks have also been effective for developing treatments: identifying drug targets (Yeh *et al*, 2012; Emig *et al*, 2013), understanding drug mechanism of action (Perez-Lopez *et al*, 2015), predicting side effects (Huang *et al*, 2013b), predicting drug–drug interactions (Huang *et al*, 2013a), and characterizing drug-regulated genes and toxicity (Kotlyar *et al*, 2012).

Biological validations

Direct experimental validation of the biological relevance of interactions is the final important step in any interactome mapping project. While complete validation of all novel interactions detected is seldom possible within the context of a single study, demonstrating that a particular interactome provides information of practical biological importance can be done by further analysis of a representative subset of interactions.

Selection of the subset of interactions to study is highly situational, and will depend largely on the nature of the proteins being investigated and the information currently available about them, as well as the size of the interactome and the specific goals of the study. For example, if studying the interactions of a protein whose mutation is known to be associated with disease, interactions which differ between the WT and mutant forms of the protein would likely be highly informative candidates for initial validation. Interactions involving members specifically associated with a given process of interest, which have not been previously demonstrated to interact, also represent a good starting point. Integrating the interactome with other datasets, combined with various predictive algorithms (as described above), is valuable in this selection process, and can help identify candidates based on a more complex range of user-defined criteria.

The specific validation experiments to be performed also vary on a case-by-case basis. Typical initial characterization experiments involve disrupting the level of individual members of an interaction pair (e.g., by gene deletion/knockdown or overexpression), and then looking for changes in the properties or function of the other member. For example, if studying the interactions of a particular receptor, one could investigate the effect of altering gene expression levels of identified interactors on downstream signaling cascades controlled by the receptor. Alternatively, effects on protein stability, protein trafficking, posttranslational modification, or responsiveness to known ligands/substrates could be probed. Investigations can be as specific as centering on molecular effects on individual proteins, or more broadly explore general phenotypic change (e.g., increased sensitivity to particular drugs, inability to grow under certain conditions). Mutational analysis of proteins can also be useful in identifying regions important for mediating and regulating interactions. For examples of proteomics screens followed by functional follow-ups, we refer the reader to several recent studies (Babu *et al*, 2012; Snider *et al*, 2013; Petschnigg *et al*, 2014).

The importance of performing these validations cannot be understated, as they provide a clear demonstration of the usefulness of any newly generated interactome in providing a solid starting point for the identification and characterization of biologically important processes. If an interactome cannot easily provide this information, then it is unlikely to be of widespread value to the scientific community, and further improvements are necessary. It is also critical to note that carrying out proper validations may represent a significant effort, and researchers must take this into account when planning and implementing any interactome mapping project, regardless of scale.

Concluding remarks

High-throughput PPI mapping and analysis enables researchers to generate data and investigate biological processes on a previously unprecedented scale (Yao *et al*, 2015). Selecting and implementing the method best suited for a particular biological question can be a significant challenge, however, one which is further complicated by the emergence of an ever increasing number of new interaction proteomics technologies. Here, we have presented the general principles of the most frequently used PPI methods and have highlighted the advantages and limitations of each, as well as provided a summary of available bioinformatics approaches and resources for use in the interpretation of interactome data. It is our hope that this review serves as a useful guide to "wet" and "dry" laboratory

© 2015 The Authors

methodologies and the analytical tools required to properly make use of these exciting approaches, and will help more scientists actively employ them in their research efforts.

Acknowledgements

We thank Dr. Ivan Plavec for valuable comments on this manuscript. The work in the Stagljar laboratory is supported by grants from the Ontario Genomics Institute, Canadian Cystic Fibrosis Foundation, Canadian Cancer Society, Pancreatic Cancer Canada and University Health Network. The Jurisica laboratory is supported in part by Ontario Research Fund (GL2-01-030), Natural Sciences Research Council (NSERC #203475), Canada Research Chair Program (CRC #203373 and #225404), Canada Foundation for Innovation (CFI #12301, #203373, #29272, #225404, #30865), US Army DOD (#W81XWH-12-1-0501), and IBM.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19: 161 – 162

Amano M, Hamaguchi T, Shohag MH, Kozawa K, Kato K, Zhang X, Yura Y, Matsuura Y, Kataoka C, Nishioka T, Kaibuchi K (2015) Kinase-interacting substrate screening is a novel method to identify kinase substrates. *J Cell Biol* 209: 895 – 912

Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, Gaulton A, Goll J, Hancock REW, Isserlin R, Jimenez RC, Kerssemakers J, Khadake J, Lynn DJ, Michaut M, O'Kelly G *et al* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* 8: 528 – 529

Audet M, Lagace M, Silversides DW, Bouvier M (2010) Protein-protein interactions monitored in cells from transgenic mice using bioluminescence resonance energy transfer. *FASEB J* 24: 2829 – 2838

Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean BDM, Burston HE, Vizeacoumar FJ, Snider J, Phanse S, Fong V, Tam YYC, Davey M, Hnatshak O, Bajaj N, Chandran S, Punna T, Christopolous C, Wong V, Yu A *et al* (2012) Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* 489: 585 – 589

Bader G, Hogue C (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2

Bader GD, Betel D, Hogue CWV (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res* 31: 248 – 250

Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12: 56 – 68

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 41: D991 – D995

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, Luga V, Przulj N, Robinson M, Suzuki H, Hayashizaki Y, Jurisica I, Wrana JL (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* 307: 1621 – 1625

Barshir R, Basha O, Eluk A, Smoly IY, Lan A, Yeger-Lotem E (2013) The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Res* 41: D841 – D844

Behrends C, Sowa ME, Gygi SP, Harper JW (2010) Network organization of the human autophagy system. *Nature* 466: 68 – 76

Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21 (Suppl 1): i38 – i46

Blasche S, Koegl M (2013) Analysis of protein-protein interactions using LUMIER assays. *Methods Mol Biol* 1064: 17 – 27

Bossi A, Lehner B (2009) Tissue specificity and the human protein interaction network. *Mol Syst Biol* 5: 260

Boute N, Jockers R, Issad T (2002) The use of resonance energy transfer in high-throughput screening: BRET versus FRET. *Trends Pharmacol Sci* 23: 351 – 354

Bovijn C, Desmet A-S, Uyttendaele I, Van Acker T, Tavernier J, Peelman F (2013) Identification of binding sites for myeloid differentiation primary response gene 88 (MyD88) and toll-like receptor 4 in MyD88 adapter-like (Mal). *J Biol Chem* 288: 12054 – 12066

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet A-S, Venkatesan K, Rual J-F, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods* 6: 91 – 97

Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ (2013) InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res* 41: D1228 – D1233

Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7: 488

Brown KR, Jurisica I (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol* 8: R95

Brown KR, Otasek D, Ali M, McGuffin MJ, Xie W, Devani B, van Toch IL, Jurisica I (2009) NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* 25: 3327 – 3329

Bushell KM, Söllner C, Schuster-Boeckler B, Bateman A, Wright GJ (2008) Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Res* 18: 622 – 630

Caligiuri M, Molz L, Liu Q, Kaplan F, Xu JP, Majeti JZ, Ramos-Kelsey R, Murthi K, Lievens S, Tavernier J, Kley N (2006) MASPIT: three-hybrid trap for quantitative proteome fingerprinting of small molecule-protein interactions in mammalian cells. *Chem Biol* 13: 711 – 722

Camargo A, Azuaje F (2007) Linking gene expression and functional network data in human heart failure. *PLoS ONE* 2: e1347

Chang DT-H, Syu Y-T, Lin P-C (2010) Predicting the protein-protein interactions using primary structures with predicted protein surface. *BMC Bioinformatics* 11 (Suppl 1): S3

Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS *et al* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43: D470 – D478

Chautard E, Fatoux-Ardore M, Ballut L, Thierry-Mieg N, Ricard-Blum S (2011) MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res* 39: D235 – D240

Chen T-C, Lin K-T, Chen C-H, Lee S-A, Lee P-Y, Liu Y-W, Kuo Y-L, Wang F-S, Lai J-M, Huang C-YF (2014) Using an in situ proximity ligation assay to systematically profile endogenous protein-protein interactions in a pathway network. *J Proteome Res* 13: 5339 – 5346

 © 2015 The Authors

Chen Y, Jiang T, Jiang R (2011) Uncover disease genes by maximizing information flow in the phenome-interactome network. *Bioinformatics* 27: i167 – i176

Chiang T, Li N, Orchard S, Kerrien S, Hermjakob H, Gentleman R, Huber W (2008) Rintact: enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics* 24: 1100 – 1101

Chin C-H, Chen S-H, Ho C-W, Ko M-T, Lin C-Y (2010) A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. *BMC Bioinformatics* 11 (Suppl 1): S25

Choi H, Larsen B, Lin Z-Y, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras A-C, Nesvizhskii AI (2011) SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* 8: 70 – 73

Chowdhary R, Tan SL, Zhang J, Karnik S, Bajic VB, Liu JS (2012) Context-specific protein network miner–an online system for exploring context-specific protein interaction networks from the literature. *PLoS ONE* 7: e34480

Ciruela F (2008) Fluorescence-based methods in the study of protein-protein interactions in living cells. *Curr Opin Biotechnol* 19: 338 – 343

Collins BC, Gillet LC, Rosenberger G, Röst HL, Vichalkovski A, Gstaiger M, Aebersold R (2013) Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat Methods* 10: 1246 – 1253

Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* 6: 439 – 450

Cooper SE, Hodimont E, Green CM (2015) A fluorescent bimolecular complementation screen reveals MAF1, RNF7 and SETD3 as PCNA-associated proteins in human cells. *Cell Cycle* 14: 2509 – 2519

Crosnier C, Bustamante LY, Bartholdson SJ, Bei AK, Theron M, Uchikawa M, Mboup S, Ndir O, Kwiatkowski DP, Duraisingh MT, Rayner JC, Wright GJ (2011) Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature* 480: 534 – 537

Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, Simonis N, Rual J-F, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M (2009) Literature-curated protein interaction datasets. *Nat Methods* 6: 39 – 46

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324 – 328

Deane CM, Salwiński Ł, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 1: 349 – 356

Deribe YL, Wild P, Chandrashaker A, Curak J, Schmidt MHH, Kalaidzidis Y, Milutinovic N, Kratchmarova I, Buerkle L, Fetchko MJ, Schmidt P, Kittanakom S, Brown KR, Jurisica I, Blagoev B, Zerial M, Stagljar I, Dikic I (2009) Regulation of epidermal growth factor receptor trafficking by lysine deacetylase HDAC6. *Sci Signal* 2: ra84

D'haeseleer P, Church GM (2004) Estimating and improving protein interaction error rates. *Proc. IEEE Comput. Syst. Bioinform. Conf.*: 216 – 223

Dingar D, Kalkat M, Chan P-K, Srikumar T, Bailey SD, Tu WB, Coyaud E, Ponzielli R, Kolyar M, Jurisica I, Huang A, Lupien M, Penn LZ, Raught B (2015) BioID identifies novel c-MYC interacting partners in cultured cells and xenograft tumors. *J Proteomics* 118: 95 – 111

van Dongen S, Abreu-Goodger C (2012) Using MCL to extract clusters from networks. *Methods Mol Biol* 804: 281 – 295

van Dongen SM (2000) *Graph clustering by flow simulation*. PhD Thesis, University of Utrecht, Utrecht

Dunham WH, Mullin M, Gingras A-C (2012) Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 12: 1576 – 1590

Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH (2011) Too many roads not taken. *Nature* 470: 163 – 165

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* 18: 529 – 536

Elefsinioti A, Saraç ÖS, Hegele A, Plake C, Hubner NC, Poser I, Sarov M, Hyman A, Mann M, Schroeder M, Stelzl U, Beyer A (2011) Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics* 10: M111.010629 A

Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, Bessarabova M (2013) Drug target prediction and repositioning using an integrated network-based approach. *PLoS ONE* 8: e60618

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86 – 90

Eyckerman S, Lemmens I, Catteeuw D, Verhee A, Vandekerckhove J, Lievens S, Tavernier J (2005) Reverse MAPPIT: screening for protein-protein interaction modifiers in mammalian cells. *Nat Methods* 2: 427 – 433

Ferro E, Trabalzini L (2013) The yeast two-hybrid and related methods as powerful tools to study plant cell signalling. *Plant Mol Biol* 83: 287 – 301

Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature* 340: 245 – 246

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E *et al* (2014) Ensembl 2014. *Nucleic Acids Res* 42: D749 – D755

Frei AP, Jeon O-Y, Kilcher S, Moest H, Henning LM, Jost C, Plückthun A, Mercer J, Aebersold R, Carreira EM, Wollscheid B (2012) Direct identification of ligand-receptor interactions on living cells and tissues. *Nat Biotechnol* 30: 997 – 1001

Frei AP, Moest H, Novy K, Wollscheid B (2013) Ligand-based receptor identification on living cells and tissues using TRICEPS. *Nat Protoc* 8: 1321 – 1336

Friedel CC, Krumsiek J, Zimmer R (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J Comput Biol* 16: 971 – 987

Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M *et al* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631 – 636

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T *et al* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141 – 147

Gingras A-C, Raught B (2012) Beyond hairballs: the use of quantitative mass spectrometry data to understand protein-protein interactions. *FEBS Lett* 586: 2723 – 2731

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. *Proc Natl Acad Sci USA* 104: 8685 – 8690

Goldberg DS, Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* 100: 4372 – 4376

Gomez SM, Noble WS, Rzhetsky A (2003) Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* 19: 1875 – 1881

© 2015 The Authors

Grossmann A, Benlasfer N, Birth P, Hegele A, Wachsmuth F, Apelt L, Stelzl U (2015) Phospho-tyrosine dependent protein-protein interaction network. *Mol Syst Biol* 11: 794

Gulati S, Balderes D, Kim C, Guo ZA, Wilcox L, Area-Gomez E, Snider J, Wolinski H, Stagljar I, Granato JT, Ruggles KV, DeGiorgis JA, Kohlwein SD, Schon EA, Sturley SL (2015) ATP-binding cassette transporters and sterol O-acyltransferases interact at membrane microdomains to modulate sterol uptake and esterification. *FASEB J* 11: 4682 – 4694

Gulbahce N, Yan H, Dricot A, Padi M, Byrdsong D, Franchi R, Lee D-S, Rozenblatt-Rosen O, Mar JC, Calderwood MA, Baldwin A, Zhao B, Santhanam B, Braun P, Simonis N, Huh K-W, Hellner K, Grace M, Chen A, Rubio R *et al* (2012) Viral perturbations of host networks reflect disease etiology. *PLoS Comput Biol* 8: e1002531

Guo Y, Li M, Pu X, Li G, Guang X, Xiong W, Li J (2010) PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. *BMC Res Notes* 3: 145

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* 36: 3025 – 3030

Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Rao S, Wang J (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* 23: 2121 – 2128

Hakes L, Pinney JW, Robertson DL, Lovell SC (2008) Protein-protein interaction networks and biology–what's the connection? *Nat Biotechnol* 26: 69 – 72

Hamdan FF, Percherancier Y, Breton B, Bouvier M (2006) Monitoring protein-protein interactions in living cells by Bioluminescence resonance Energy Transfer (BRET). *Curr Prot Neurosci* 34: 5.23.1 – 5.23.20

Hamdi A, Colas P (2012) Yeast two-hybrid methods and their applications in drug discovery. *Trends Pharmacol Sci* 33: 109 – 118

Hart GT, Lee I, Marcotte ER (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* 8: 236

Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar V-N, Bezginov A, Clark GW, Wu GC, Wodak SJ *et al* (2012) A census of human soluble protein complexes. *Cell* 150: 1068 – 1081

Hirsh E, Sharan R (2007) Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics* 23: e170 – e176

Hu C-D, Chinenov Y, Kerppola TK (2002) Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Mol Cell* 9: 789 – 798

Huang H, Bader JS (2009) Precision and recall estimates for two-hybrid screens. *Bioinformatics* 25: 372 – 378

Huang J, Niu C, Green CD, Yang L, Mei H, Han J-DJ (2013a) Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Comput Biol* 9: e1002998

Huang L-C, Wu X, Chen JY (2013b) Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics* 13: 313 – 324

Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 (Suppl 1): S233 – S240

Ideker T, Sharan R (2008) Protein networks in disease. *Genome Res* 18: 644 – 652

Ivanic J, Yu X, Wallqvist A, Reifman J (2009) Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS ONE* 4: e5815

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302: 449 – 453

Jung SH, Hyun B, Jang W-H, Hur H-Y, Han D-S (2010) Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics* 26: 385 – 391

Kerppola TK (2008) Bimolecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells. *Annu Rev Biophys* 37: 465 – 487

Kerr JS, Wright GJ (2012) Avidity-based extracellular interaction screening (AVEXIS) for the scalable detection of low-affinity extracellular receptor-ligand interactions. *J Vis Exp* 61: e3881

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B *et al* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 40: D841 – D846

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J, Cerami E *et al* (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol* 5: 44

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M *et al* (2009) Human protein reference database–2009 update. *Nucleic Acids Res* 37: D767 – D772

Kim DI, Birendra KC, Zhu W, Motamedchaboki K, Doye V, Roux KJ (2014) Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc Natl Acad Sci USA* 111: E2453 – E2461

King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* 20: 3013 – 3020

Kocan M, See HB, Seeber RM, Eidne KA, Pfleger KDG (2008) Demonstration of improvements to the bioluminescence resonance energy transfer (BRET) technology for the monitoring of G protein-coupled receptors in live cells. *J Biomol Screen* 13: 888 – 898

Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82: 949 – 958

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res* 43: D1113 – D1116

Koos B, Andersson L, Clausson C-M, Grannas K, Klaesson A, Cane G, Söderberg O (2014) Analysis of protein interactions in situ by proximity ligation assays. *Curr Top Microbiol Immunol* 377: 111 – 126

Kotlyar M, Fortney K, Jurisica I (2012) Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* 57: 499 – 507

Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafaee F, Broackes-Carter F, Petschnigg J, Mills GB, Jurisicova A, Stagljar I, Maestro R, Jurisica I (2015) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods* 12: 79 – 84

Kotlyar M, Pastrello C, Sheahan N, Jurisica I (2016) Integrated Interactions Database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res* doi:10.1093/nar/gkv1115

 © 2015 The Authors

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B *et al* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440: 637 – 643

Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 25: 309 – 316

Lam MHY, Snider J, Rehal M, Wong V, Aboualizadeh F, Drecun L, Wong O, Jubran B, Li M, Ali M, Jessulat M, Deineko V, Miller R, Lee ME, Park H-O, Davidson A, Babu M, Stagljar I (2015) A comprehensive membrane interactome mapping of Sho1p reveals Fps1p as a novel key player in the regulation of the HOG pathway in *S. cerevisiae*. *J Mol Biol* 427: 2088 – 2103

Lambert J-P, Tucholska M, Go C, Knight JDR, Gingras A-C (2015) Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *J Proteomics* 118: 81 – 94

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011a) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 21: 1109 – 1121

Lee ME, Singh K, Snider J, Shenoy A, Paumi CM, Stagljar I, Park H-O (2011b) The Rho1 GTPase acts together with a vacuolar glutathione S-conjugate transporter to protect yeast cells from oxidative stress. *Genetics* 188: 859 – 870

Lemmens I, Lievens S, Tavernier J (2015) MAPPIT, a mammalian two-hybrid method for in-cell detection of protein-protein interactions. *Methods Mol Biol* 1278: 447 – 455

Leung HCM, Xiang Q, Yiu SM, Chin FYL (2009) Predicting protein complexes from PPI data: a core-attachment approach. *J Comput Biol* 16: 133 – 144

Li X, Wu M, Kwoh C-K, Ng S-K (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genom* 11 (Suppl 1): S3

Li X-L, Foo C-S, Ng S-K (2007) Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. *Comput Syst Bioinformatics Conf* 6: 157 – 168

Li X-L, Tan S-H, Foo C-S, Ng S-K (2005) Interaction graph mining for protein complexes using local clique merging. *Genome Inform* 16: 260 – 269

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40: D857 – D861

Lievens S, Gerlo S, Lemmens I, De Clercq DJH, Risseeuw MDP, Vanderroost N, De Smet A-S, Ruyssinck E, Chevet E, Van Calenbergh S, Tavernier J (2014) Kinase Substrate Sensor (KISS), a mammalian in situ protein interaction sensor. *Mol Cell Proteomics* 13: 3332 – 3342

Lievens S, Peelman F, De Bosscher K, Lemmens I, Tavernier J (2011) MAPPIT: a protein interaction toolbox built on insights in cytokine receptor signaling. *Cytokine Growth Factor Rev* 22: 321 – 329

Lievens S, Vanderroost N, Van der Heyden J, Gesellchen V, Vidal M, Tavernier J (2009) Array MAPPIT: high-throughput interactome analysis in mammalian cells. *J Proteome Res* 8: 877 – 886

Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125: 801 – 814

Lin C-C, Hsiang J-T, Wu C-Y, Oyang Y-J, Juan H-F, Huang H-C (2010) Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst Biol* 4: 138

Liu G, Li J, Wong L (2008) Assessing and predicting protein interactions using both local and global network topological metrics. *Genome Inform* 21: 138 – 149

Liu G, Wong L, Chua HN (2009) Complex discovery from weighted PPI networks. *Bioinformatics* 25: 1891 – 1897

Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35: D198 – D201

Lopes TJS, Schaefer M, Shoemaker J, Matsuoka Y, Fontaine J-F, Neumann G, Andrade-Navarro MA, Kawaoka Y, Kitano H (2011) Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27: 2414 – 2421

Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* 49: 350 – 364

Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci USA* 100: 5772 – 5777

Ma L, Yang F, Zheng J (2014) Application of fluorescence resonance energy transfer in protein studies. *J Mol Struct* 1077: 87 – 100

Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 39: D52 – D57

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011: bar009

Mandić M, Drinovec L, Glisic S, Veljkovic N, Nøhr J, Vrecl M (2014) Demonstration of a direct Interaction between β2-adrenergic receptor and insulin receptor by BRET and bioinformatics. *PLoS ONE* 9: e112664

Martin S, Roe D, Faulon J-L (2005) Predicting protein-protein interactions using signature products. *Bioinformatics* 21: 218 – 226

Martin S, Sollner C, Charoensawan V, Adryan B, Thisse B, Thisse C, Teichmann S, Wright GJ (2010) Construction of a large extracellular protein interaction network and its resolution by spatiotemporal expression profiling. *Mol Cell Proteomics* 9: 2654 – 2665

Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY, Halim VA, Bagshaw RD, Hubner NC, Al-Hakim A, Bouchard A, Faubert D, Fermin D, Dunham WH, Goudreault M, Lin Z-Y *et al* (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 10: 730 – 736

Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S (2005) Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci USA* 102: 12123 – 12128

Miller KE, Kim Y, Huh W-K, Park H-O (2015) Bimolecular fluorescence complementation (BiFC) analysis: advances and recent applications for genome-wide interaction studies. *J Mol Biol* 427: 2039 – 2055

Nanni L, Lumini A (2006) An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics* 22: 1207 – 1210

Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26: 1057 – 1063

Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 9: 471 – 472

Nguyen TP, Ho TB (2006) Discovering signal transduction networks using signaling domain-domain interactions. *Genome Inform* 17: 35 – 45

Orchard S (2012) Molecular interaction databases. *Proteomics* 12: 1656 – 1662

© 2015 The Authors

Orchard S, Kerrien S (2010) Molecular interactions and data standardisation. *Methods Mol Biol* 604: 309 – 318

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FSL, Brinkman F, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock REW, Hancock R, Hannick LI, Jurisica I *et al* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* 9: 345 – 350

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin A-C, Superti-Furga G, Greenblatt J *et al* (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol* 25: 894 – 898

Ourfali O, Shlomi T, Ideker T, Ruppin E, Sharan R (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* 23: i359 – i366

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* 1: 93 – 108

Ozawa Y, Saito R, Fujimori S, Kashima H, Ishizaka M, Yanagawa H, Miyamoto-Sato E, Tomita M (2010) Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions. *BMC Bioinformatics* 11: 350

Paumi CM, Menendez J, Arnoldo A, Engels K, Iyer KR, Thaminy S, Georgiev O, Barral Y, Michaelis S, Stagljar I (2007) Mapping protein-protein interactions for the yeast ABC transporter Ycf1p by integrated split-ubiquitin membrane yeast two-hybrid analysis. *Mol Cell* 26: 15 – 25

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96: 4285 – 4288

Perez-Lopez ÁR, Szalay KZ, Türei D, Módos D, Lenti K, Korcsmáros T, Csermely P (2015) Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. *Sci Rep* 5: 10182

Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C (2010) Transient protein-protein interactions: structural, functional, and network properties. *Structure* 18: 1233 – 1243

Petschnigg J, Groisman B, Kotlyar M, Taipale M, Zheng Y, Kurat CF, Sayad A, Sierra JR, Mattiazzi Usaj M, Snider J, Nachman A, Krykbaeva I, Tsao M-S, Moffat J, Pawson T, Lindquist S, Jurisica I, Stagljar I (2014) The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nat Methods* 11: 585 – 592

Petschnigg J, Wong V, Snider J, Stagljar I (2012) Investigation of membrane protein interactions using the split-ubiquitin membrane yeast two-hybrid system. *Methods Mol Biol* 812: 225 – 244

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 40: D130 – D135

Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* 7: 944 – 960

Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Häuser R, Siszler G, Wuchty S, Emili A, Babu M, Aloy P, Pieper R, Uetz P (2014) The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* 32: 285 – 290

Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* 23: 951 – 959

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis A-R *et al* (2014) A proteome-scale map of the human interactome network. *Cell* 159: 1212 – 1226

Roux KJ, Kim DI, Raida M, Burke B (2012) A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* 196: 801 – 810

Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M (2009) Exploiting amino acid composition for predicting protein-protein interactions. *PLoS ONE* 4: e7813

Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y, Kovács IA, Kamburov A, Krykbaeva I, Lam MH, Tucker G, Khurana V, Sharma A, Liu Y-Y, Yachie N, Zhong Q *et al* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161: 647 – 660

Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, Vidal M (2013) Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* 23: 649 – 657

Saito R, Suzuki H, Hayashizaki Y (2002) Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res* 30: 1163 – 1168

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32: D449 – D451

Sanderson CM (2008) A new way to explore the world of extracellular protein interactions. *Genome Res* 18: 517 – 520

Sauvageau E, Rochdi MD, Oueslati M, Hamdan FF, Percherancier Y, Simpson JC, Pepperkok R, Bouvier M (2014) CNIH4 interacts with newly synthesized GPCR and controls their export from the endoplasmic reticulum. *Traffic* 15: 383 – 400

Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. *Nature* 461: 218 – 223

Schulze WX, Deng L, Mann M (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol Syst Biol* 1(2005): 0008

Schwartz AS, Yu J, Gardenour KR, Finley RL, Ideker T (2009) Cost-effective strategies for completing the interactome. *Nat Methods* 6: 55 – 61

Sharan R, Ideker T, Kelley B, Shamir R, Karp RM (2005) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol* 12: 835 – 846

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H (2007) Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA* 104: 4337 – 4341

Simonis N, Rual J-F, Carvunis A-R, Tasan M, Lemmens I, Hirozane-Kishikawa T, Hao T, Sahalie JM, Venkatesan K, Gebreab F, Cevik S, Klitgord N, Fan C, Braun P, Li N, Ayivi-Guedehoussou N, Dann E, Bertin N, Szeto D, Dricot A *et al* (2009) Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat Methods* 6: 47 – 54

Singhal M, Resat H (2007) A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics* 8: 199

Sinha R, Kundrotas PJ, Vakser IA (2010) Docking by structural similarity at protein-protein interfaces. *Proteins* 78: 3235 – 3241

Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, Veldboer J, Zemojtel T, Robinson PN (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* 30: 3215 – 3222

 © 2015 The Authors

Snider J, Hanif A, Lee ME, Jin K, Yu AR, Graham C, Chuk M, Damjanovic D, Wierzbicka M, Tang P, Balderes D, Wong V, Jessulat M, Darowski KD, San Luis B-J, Shevelev I, Sturley SL, Boone C, Greenblatt JF, Zhang Z *et al* (2013) Mapping the functional yeast ABC transporter interactome. *Nat Chem Biol* 9: 565 – 572

Snider J, Kittanakom S, Damjanovic D, Curak J, Wong V, Stagljar I (2010) Detecting interactions with membrane proteins using a membrane two-hybrid assay in yeast. *Nat Protoc* 5: 1281 – 1293

Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol* 311: 681 – 692

Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein-protein interaction data? *J Mol Biol* 327: 919 – 923

Srihari S, Leong HW (2013) A survey of computational methods for protein complex prediction from protein interaction networks. *J Bioinform Comput Biol* 11: 1230002

Srihari S, Ning K, Leong HW (2010) MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. *BMC Bioinformatics* 11: 504

Stagljar I, Korostensky C, Johnsson N, te Heesen S (1998) A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proc Natl Acad Sci USA* 95: 5187 – 5192

Stasi M, De Luca M, Bucci C (2015) Two-hybrid-based systems: powerful tools for investigation of membrane traffic machineries. *J Biotechnol* 202: 105 – 117

Su G, Morris JH, Demchak B, Bader GD (2014) Biological network exploration with cytoscape 3. *Curr Protoc Bioinformatics* 47: 8.13.1 – 8.13.24

Sun Y, Gallagher-Jones M, Barker C, Wright GJ (2012) A benchmarked protein microarray-based platform for the identification of novel low-affinity extracellular protein interactions. *Anal Biochem* 424: 45 – 53

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: D447 – D452

Taipale M, Tucker G, Peng J, Krykbaeva I, Lin Z-Y, Larsen B, Choi H, Berger B, Gingras A-C, Lindquist S (2014) A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell* 158: 434 – 448

Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ (2014) Navigating the global protein-protein interaction landscape using iRefWeb. *Methods Mol Biol* 1091: 315 – 331

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T *et al* (2015) Tissue-based map of the human proteome. *Science* 347: 1260419

Ulrichts P, Lemmens I, Lavens D, Beyaert R, Tavernier J (2009) MAPPIT (mammalian protein-protein interaction trap) analysis of early steps in toll-like receptor signalling. *Methods Mol Biol* 517: 133 – 144

Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 6: e1000641

Varjosalo M, Sacco R, Stukalov A, van Drogen A, Planyavsky M, Hauri S, Aebersold R, Bennett KL, Colinge J, Gstaiger M, Superti-Furga G (2013) Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat Methods* 10: 307 – 314

Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K-I, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet A-S, Dann E, Smolyar A *et al* (2009) An empirical framework for binary interactome mapping. *Nat Methods* 6: 83 – 90

Vidal M, Cusick ME, Barabási A-L (2011) Interactome networks and human disease. *Cell* 144: 986 – 998

Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10: 99

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287: 116 – 122

Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ, Koller D (2009) A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. *Mol Cell Proteomics* 8: 1361 – 1381

Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23: 1274 – 1281

Wang X, Huang L (2008) Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. *Mol Cell Proteomics* 7: 46 – 57

Wang X, Huang L (2014) Defining dynamic protein interactions using SILAC-based quantitative mass spectrometry. *Methods Mol Biol* 1188: 191 – 205

Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* 30: 159 – 164

Wang Z, Clark NR, Ma'ayan A (2015) Dynamics of the discovery process of protein-protein interactions from low content studies. *BMC Syst Biol* 9: 26

Weimann M, Grossmann A, Woodsmith J, Özkan Z, Birth P, Meierhofer D, Benlasfer N, Valovka T, Timmermann B, Wanker EE, Sauer S, Stelzl U (2013) A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat Methods* 10: 339 – 342

Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, Rümmele P, Jahnke B, Hentrich V, Rückert F, Niedergethmann M, Weichert W, Bahra M, Schlitt HJ, Settmacher U, Friess H, Büchler M, Saeger H-D, Schroeder M, Pilarsky C *et al* (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 8: e1002511

Wojcik J, Schächter V (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17 (Suppl 1): S296 – S305

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z *et al* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318: 1108 – 1113

Woodsmith J, Stelzl U (2014) Studying post-translational modifications with protein interaction networks. *Curr Opin Struct Biol* 24: 34 – 44

Wu M, Li X, Kwoh C-K, Ng S-K (2009) A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinformatics* 10: 169

Xiao Y, Xu C, Xu L, Guan J, Ping Y, Fan H, Li Y, Zhao H, Li X (2012) Systematic identification of common functional modules related to heart failure with different etiologies. *Gene* 499: 332 – 338

Xie Q, Soutto M, Xu X, Zhang YJC (2011) Bioluminescence resonance energy transfer (BRET) imaging in plant seedlings and mammalian cells. *Methods Mol Biol* 680: 3 – 28

© 2015 The Authors

Xu G, Barrios-Rodiles M, Jerkic M, Turinsky AL, Nadon R, Vera S, Voulgaraki D, Wrana JL, Toporsian M, Letarte M (2014) Novel protein interactions with endoglin and activin receptor-like kinase 1: potential role in vascular networks. *Mol Cell Proteomics* 13: 489 – 502

Yao Z, Petschnigg J, Ketteler R, Stagljar I (2015) Application guide for omics approaches to cell signaling. *Nat Chem Biol* 11: 387 – 397

Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, Auluck PK, Geddie ML, Valastyan JS, Karger DR, Lindquist S, Fraenkel E (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* 41: 316 – 323

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci USA* 101: 5934 – 5939

Yeh S-H, Yeh H-Y, Soo V-W (2012) A network flow approach to predict drug targets from microarray data, disease genes and interactome network – case study on prostate cancer. *J Clin Bioinforma* 2: 1

Yoon D, Kim H, Suh-Kim H, Park RW, Lee K (2011) Differentially co-expressed interacting protein pairs discriminate samples under distinct stages of HIV type 1 infection. *BMC Syst Biol* 5 (Suppl 2): S1

Yu C-Y, Chou L-C, Chang DT-H (2010) Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* 11: 167

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J-F, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C *et al* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322: 104 – 110

Yu H, Lin C-C, Li Y-Y, Zhao Z (2013) Dynamic protein interaction modules in human hepatocellular carcinoma progression. *BMC Syst Biol* 7(Suppl 5): S2

Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han J-DJ, Bertin N, Chung S, Vidal M, Gerstein M (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 14: 1107 – 1118

Yu X, Ivanic J, Wallqvist A, Reifman J (2009) A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput Biol* 5: e1000515

Zaki N, Lazarova-Molnar S, El-Hajj W, Campbell P (2009) Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics* 10: 150

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B (2012a) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490: 556 – 560

Zhang X, Yang H, Gong B, Jiang C, Yang L (2012b) Combined gene expression and protein interaction analysis of dynamic modularity in glioma prognosis. *J Neurooncol* 107: 281 – 288

Zhang X-E, Cui Z, Wang D (2015) Sensing of biomolecular interactions using fluorescence complementing systems in living cells. *Biosens Bioelectron* 76: 243 – 250

Zhong Q, Simonis N, Li Q-R, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, Yildirim MA, Yan H, Dricot A, Szeto D, Lin C, Hao T, Fan C, Milstein S, Dupuy D *et al* (2009) Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* 5: 321

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Biophysical Reviews (2022) 14:1341–1348
https://doi.org/10.1007/s12551-022-01032-7

REVIEW

Protein–protein interaction prediction methods: from docking-based to AI-based approaches

Yuko Tsuchiya[1] · **Yu Yamamori**[1] · **Kentaro Tomii**[1]

Received: 30 September 2022 / Accepted: 30 November 2022 / Published online: 17 December 2022
© The Author(s) 2022

Abstract
Protein–protein interactions (PPIs), such as protein–protein inhibitor, antibody–antigen complex, and supercomplexes play diverse and important roles in cells. Recent advances in structural analysis methods, including cryo-EM, for the determination of protein complex structures are remarkable. Nevertheless, much room remains for improvement and utilization of computational methods to predict PPIs because of the large number and great diversity of unresolved complex structures. This review introduces a wide array of computational methods, including our own, for estimating PPIs including antibody–antigen interactions, offering both historical and forward-looking perspectives.

Keywords Protein–protein interaction · Antibody–antigen interaction · Epitope prediction · Docking simulation · AlphaFold · Profile–profile comparison

Introduction

Protein–protein interactions (PPIs) play fundamentally important roles in cellular functions and biological processes, and structural understanding of the PPIs is important for the elucidation of those functions (Jones and Thornton 1995, 1996). In 2001, the Critical Assessment of PRediction of Interactions (CAPRI, 2022) began as a community-wide experiment designed to assess methods for predicting PPIs based on the estimation of PPIs for previously solved structures of protein complexes. The latest experiment (Round 54) was conducted during May–August in 2022 (CAPRI Round 54, 2022). A recent report of a CAPRI experiment (Lensink et al. 2020) indicated that the increase of structures of protein complexes deposited in the Protein Data Bank (PDB) enables their use as structural complex templates for predicting the structures of other protein complexes, particularly in homo protein–protein docking. That finding implies that classical docking may no longer be strictly necessary for the prediction of homo PPIs. Nevertheless, these methods have not yet been successfully applied to the prediction of hetero PPIs, including antibody–antigen interactions, and this area of research has room for improvement. In this review, we place our focus on computational docking-based approaches and AI-based approaches, introducing a wide array of methods, including our own, for PPI prediction.

Protein–protein docking

Traditional protein–protein docking methods have been of central importance for sampling the conformational space of protein complexes (Smith and Sternberg 2002). In the last 10 years, sophisticated high-precision docking methods such as HADDOCK (van Zundert et al. 2015), ClusPro (Desta et al. 2020), ZDOCK (Pierce et al. 2014), and LightDock (Jiménez-García et al. 2018) have been developed and continually improved. To sample conformations more efficiently, they have used not only three-dimensional structural information in both modeling and scoring steps, but also evolutionary information and information obtained from experiments of several types (van Noort et al. 2021). Evolutionary information has been used for the prediction of PPIs based on the idea that mutations of the residues involved in the interaction on a protein engender mutations of the interface residues on the partner protein (Jothi et al. 2006). The InterEvDock docking pipeline integrates a coarse-grained potential accounting for interface coevolution

✉ Kentaro Tomii
k-tomii@aist.go.jp

[1] Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-Ku, Tokyo 135-0064, Japan

based on multiple sequence alignments (MSAs) of paired proteins (Yu et al. 2016). For many targets in recent CAPRI experiments, this method predicted several high-quality models (Lensink et al. 2020).

Professor Nakamura and his colleagues, including one of the authors of the present review (Y.T.), participated in CAPRI about 10 years ago (Fleishman et al. 2011; Moretti et al. 2013; Lensink et al. 2014). For each CAPRI experiment target, the docking of two component (mainly unbound) protein structures was performed to generate many complex models, and the position and relative geometry of the predicted interfaces were evaluated based on the scoring function. We constructed docking methods with and without evolutional tracing using conservation information, where the conservation information contributed to model selections particularly for the targets of enzyme and signal transduction proteins (Kanamori et al. 2007).

Recently, data of low-resolution shapes obtained by small-angle scattering methods such as SAXS, were also used for the filtering of docked models in several docking methods (van Noort et al. 2021), e.g., pyDockSAXS (Jiménez-García et al. 2015) and HADDOCK (van Zundert et al. 2015).

Scoring models

After sampling complex models, the most "appropriate" model(s) must usually be selected using some method such as scoring function, clustering, or consensus (Lensink and Wodak 2010). In CAPRI, the "scorer" performance is measured and reported as part of the competition (Lensink and Wodak 2010). Predicted models for each target were pooled and provided for the scoring experiment. Based on the idea that the accumulation of complementary interactions on local (residue or atomic level) surface regions forms the PPI between two proteins and strengthens their interaction, we specifically examined the complementarities of physicochemical properties such as electrostatic potential and hydrophobicity, and shapes of the surfaces to assess PPIs on the molecular surfaces of proteins (Tsuchiya et al. 2006a, b).

Machine-learning (ML) and AI-based scoring methods have been developed in the past few years (Lensink et al. 2020). The ML-based iScore is a linearly combined score of GraphRank with HADDOCK energetic terms (Geng et al. 2020). The GraphRank score is based on a support vector machine (SVM) classification between the interface graphs of native and non-native protein–protein interfaces, where the graph consists of nodes (interface residues with evolutionary conservation scores) and edges (their contacts) (Geng et al. 2020). The iScore separated native from non-native interfaces much better than the case of individual usage of the GraphRank or HADDOCK score, when the GraphRank contribution is higher than the energetic terms. Das et al. also developed an SVM classification-based scoring scheme by particularly addressing differences between native and non-native protein–protein interface features such as binding energy (ΔG), frequencies of hydrogen bonds and salt-bridges, and accessible and buried surface areas (Das and Chakrabarti 2021). Their findings demonstrated the accessible surface area as the most distinguishable feature. In native interfaces, the buried surface areas of Phe, Tyr, and Ile are markedly high whereas that of Lys is low, and hydrogen bonds in Arg with Asp or Glu are more important (Das and Chakrabarti 2021). Scoring based on these features led to higher accuracy than that provided by PatchDock scores (Schneidman-Duhovny et al. 2005). Additionally, we have developed the Protein Interface Analysis using COvarying signals (PIACO 2019), a method for identifying biological interfaces in protein crystal structures. PIACO is a statistical classifier that uses covariation calculated from MSAs as input (Fukasawa and Tomii 2019).

Another recently developed AI-based scoring method, known as DOVE, is based on the use of three-dimensional (3D) convolutional neural networks (Wang et al. 2020). It first maps the inputted decoy structures into a 3D grid, then scans and examines the protein–protein interfaces in terms of inter-atom interaction patterns and their energetic contributions. Finally, DOVE judges whether the decoy structure is close to the native structure, or not.

Hot spots

In PPI prediction, a hot spot is a prominent feature that allows one to distinguish between native and non-native interface models (DeLano 2002). Typically, a hot spot is a small set of residues that contribute greatly to PPI formation (Thorn and Bogan 2001). From the converse perspective, blocking hot spot residues might engender the disruption of the PPIs (Lu et al. 2020). Hot spots are fundamentally identified based on changes in binding free energy (ΔΔG) obtained by alanine scanning mutagenesis, whereas computational methods of the accurate predictions of hot spots have also been developed (Ovek et al. 2022). Ovek et al. introduce two interesting examples in their recent review (Ovek et al. 2022): disease-causing single mutations are likely to occur on hot spots (Ozdemir et al. 2018) and disease-causing PPI can be disrupted by a small molecule that binds to the target hot spots on the PPI (Lim et al. 2019). These findings suggest that hot spot information might contribute to the discovery of drugs that disrupt the disease-causing PPI (Ovek et al. 2022).

When we participated in the CAPRI experiment, several groups tried to predict hot spots to select near native models. Of those, Fernandez-Recio's group predicted hot spots using the normalized interface propensity values derived

from rigid-body docking, with electrostatics and desolvation scoring, which required no structural information of protein–protein complexes (Grosdidier and Fernández-Recio 2008). The same group recently constructed a method to detect protein–protein inhibitor (small molecule) binding sites by integrating molecular dynamics simulations for the generation of transient cavities on the interface with hot-spot predictions (Rosell and Fernández-Recio 2020). It generated and selected the transient cavities that were similar to known inhibitor binding sites (Rosell and Fernández-Recio 2020).

Template-based method

As we mentioned, protein–protein docking has been centrally important to PPI prediction. However, along with the increase of solved complex structures, it has been found to be more helpful to use them as templates for PPI prediction. As mentioned earlier (Lensink et al. 2020), found that the use of solved complex structures as templates is particularly appropriate for homo PPI prediction because of the abundance of such homo PPI structures within the PDB.

As an example, we introduce work from our own group in which we have developed a profile–profile comparison method: Fold Recognition TEchnique (FORTE) (Tomii and Akiyama 2004). The method was developed originally for predicting protein structures. We recognized that detecting and using complex templates is effective for PPI prediction (even for hetero PPI in some cases) through participation in past CAPRI-CASP assembly prediction experiments (Lensink et al. 2016, 2018; Nakamura et al. 2018). We also found that FORTE is helpful for selecting an appropriate form of the protein complex, even for cases with multiple forms of complexes observed among homologous proteins of a target (Nakamura et al. 2018). Indeed, as we mentioned below, most of AI-based approaches which utilize information of templates are powerful for PPI, except in the case of antibody–antigen interaction prediction (Ambrosetti et al. 2020).

As a general rule, the prediction capability for generating accurate target structures depends on the existence of available templates in the PDB (Lafita et al. 2018). In CAPRI-CASP12, according to the assessors' definition, there were three levels of difficulty for estimating target complexes—EASY, MEDIUM, and HARD. We showed that template-based models based on profile–profile comparison methods are useful for predicting protein complexes, even for MEDIUM/HARD targets (Nakamura et al. 2018; Lafita et al. 2018; Lensink et al. 2018). This finding implies that PPIs also tend to be conserved and/or limited for many cases, although it is considered that complex structures are often not conserved during evolution (Poupon and Janin 2010).

Earlier attempts of application of AlphaFold2 to PPI prediction

When the great achievement of AlphaFold2 (AF2) for the modeling of monomeric structures of protein in CASP14 was announced, it was assumed that AF2 would be useful for the modeling of protein–protein complexes, even though it was originally trained on individual protein chains (Jumper et al. 2021). RoseTTaFold, the first attempt to replicate AF2 before the release of the AF2 codes, confirmed that deep neural network model trained for monomer structure prediction can predict protein–protein complex models with some degree of accuracy (Baek et al. 2021). They used the pseudo-multimer sequence, two or more sequences with a gap separating them, (and templates) as input instead of a monomer sequence. AF2 and RoseTTaFold have since been used for predicting PPIs in proteomes of human (Burke et al. 2022) and yeast (Humphreys et al. 2021).

Similar tricks to that developed in RoseTTaFold were also applied for AF2, such as AF2-Gap (in ColabFold (Mirdita et al. 2022)) and AF2-Linker (Moriwaki 2021; Evans et al. 2021). Other groups applied similar protocols for protein–peptide docking (Ko and Lee 2021; Lei et al. 2021; Tsaban et al. 2022). All of these used the pseudo-multimer sequence linking the peptide to the protein via poly-glycine as input. The protocol that was tested for the benchmark set shows higher performance than the previous state-of-the-art method (Tsaban et al. 2022). FoldDock, another sophisticated AF2-based multimer prediction method using a paired MSA, was developed and tested for a large benchmark set of heterodimers (Bryant et al. 2022a). These results underscore the superiority of AF2-based methods to other docking methods. Although the method described above is powerful for dimers (or trimer, perhaps), they might have limitations because of the difficulty in preparing good MSAs of larger complexes as their input. In contrast, MolPC is an ambitious attempt for using AF2 to build a larger complex (more than 10 chains) combining the predicted subcomponents using Monte Carlo tree search (Bryant et al. 2022b). The AF2Complex uses the MSAs of each chain by padding the gaps and templates as inputs and by generating complex models by the AF2 deep learning model after multiple recycling steps (Gao et al. 2022).

A combination of AF2 and an existing docking method was also tested (Ghani et al. 2021). The AF2-ClusPro method was adopted for the protein–protein docking benchmark 5.0 (BM5) (Vreven et al. 2015). Then we selected 17 heteromers which appeared in the PDB after May 2020 (the last date of structures utilized in the AF2 training set). The systematic benchmark of 152 diverse

heterodimers from BM5.5 data set (Vreven et al. 2015; Guest et al. 2021) was expanded by adding more antibody–antigen cases to BM5, with results revealing a significant ability of AF2-based methods for all categories of modeling of complexes (except for modeling of antibody–antigen complexes) (Yin et al. 2022), approaches represented early attempts for AF2-based multimer prediction done mainly during the period between the release of the original AF2 (ver. 2.0.0) and that of the specifically trained AF2 for multimer prediction (ver. 2.1.0).

Multimetric version of AF2 and recent attempts to go beyond AF2

The multimeric version of AF2, AF2-Multimer, appeared recently (Evans et al. 2021). Benchmarking results obtained for the selected 17 heteromer and a large data set composed of recently released 4433 complexes was compared with AF2-Gap, AF2-linker and AF2-ClusPro, and it showed improvement in multimer prediction accuracy compared with those input-adjustment AF2-based multimer predictions. It was also demonstrated that the performance is generally higher for homomeric prediction than for heteromeric prediction, and that it is poor for prediction of the antibody–antigen complex. The latter result was also clarified by the benchmark of another group (Yin et al. 2022).

More recently, two avenues of approaches were used to go beyond the AF2 and AF2-Multimer which involved the development of replication of AF2 with a larger training dataset and development of the predictor using the protein language model (pLM) instead of MSA. Among the follow-up replicative studies of AF2, such as OpenFold (Ahdritz et al. 2022), Uni-Fold (Li et al. 2022a, b), MegaFold (Liu et al. 2022), and HelixFold (Wang et al. 2022), it was only Uni-Fold and Uni-Fold-symmetry (Li et al. 2022b) that could succeed not only in monomer prediction but also in multimer prediction upon using the original trained parameters or protocols. No third-party group has yet published a large benchmark result. Actually, pLM-based predictors such as OmegaFold (Wu et al. 2022), ESMFold (Lin et al. 2022), and IgFold (Ruffolo et al. 2022) are anticipated as the next breakthroughs in structure prediction because they omit the construction of MSA, which is crucially important for performance, and which is the most time-consuming part of AF2. In spite of these expectations, no MSA-free method has achieved performance equal to that of AF2, with no explicit implementation of multimer modeling with the exception that IgFold was built specifically for antibody modeling.

Although no explicit implementation of multimer treatment in OmegaFold has been described, it is relatively straightforward to use pseudo-multiple sequence inputs as an AF2-linker which we designate as OmegaFold-linker. Figure 1 presents multimer modeling results in terms of the DockQ score (Basu and Wallner 2016) using methods developed post AF2 such as OmegaFold-linker, AF2-Multimer (v2.1.1) with parameters released on Nov. 2021, and AF2-Multimer (v.2.2.0) with parameters released on Mar. 2022. In the construction of Fig. 1, AF2 was run with "max_template_date = 2020–10–01". The targets are taken from those of CASP14 (not included in the training set of AF2 and AF2-multimer, but two of these structures (PDB IDs: 6N64 and 6YA2) of them are released before 2020–10–01). They are all dimers, and the target IDs are T1032 (PDB ID: 6N64), T1038 (PDB ID: 6YA2), T1054 (PDB ID: 6V4V), T1078 (PDB ID: 7CWP), H1045 (PDB ID: 6XOD), and H1065 (PDB ID: 7M5F). For OmegaFold-linker, due to the memory-limitation, T1032 with more than 500 residues

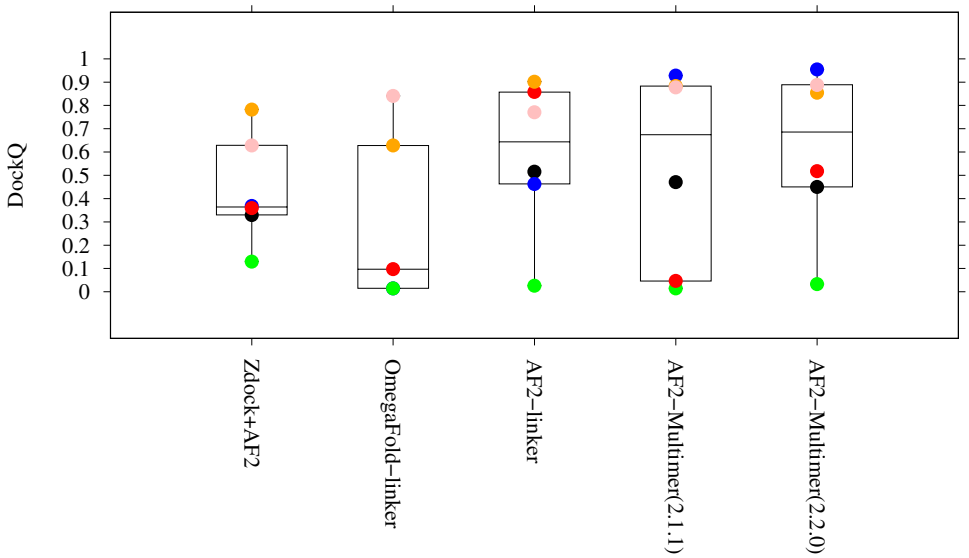


Fig. 1 Distributions of DockQ scores as boxplots for different modeling methods: ZDOCK with AF2 models, OmegaFold-linker, AF2-linker, AF2-Multimer (v.2.1.1), and AF2-Multimer (v2.2.0). Six targets (T1032 (shown in black dot), T1038 (blue), T1054 (green), T1078 (red), H1045 (orange), and H1065 (pink)) are taken from CASP14 (see text). For the case of OmegaFold-linker, T1032 was omitted because its amino acid length is beyond the limitation of OmegaFold

was omitted. For Omega-linker and AF2-linker, the input sequences were connected via 21 length poly-GLY-GLY-SER linker according to the previous research (Evans et al. 2021). For comparison, DockQ scores of docking structures calculated using ZDOCK (Pierce et al. 2014) are also shown. As the initial monomer structures of ZDOCK, the same models predicted by AF2-Multimer (v.2.2.0) were used. Basically, the prediction difficulty depends on targets. For instance, all methods except ZDOCK show 0.77 or more for H1065, while all methods show DockQ < 0.04 for T1054. These trends, except for T1038 (6YA2) whose structure was released before 2020–10–01, are generally similar to those observed in CASP14. Although, this may imply that the prediction difficulty of PPI depends on the availability of "good" templates, the results accord with those reported from benchmark research (Evans et al. 2021; Bryant et al. 2022a): AF2-Multimer shows the highest performance, and AF2-based methods using parameters trained for monomers are still comparable in some cases.

Antibody–antigen interaction

The prediction of PPI between antibody and antigen proteins is not an easy task because of the flexibility of the antibody's hypervariable loops, particularly the complementary determining loop 3 in the heavy chain (CDR-H3 loop). Four high-precision docking software suites ClusPro (Brenke et al. 2012), LightDock (Jiménez-García et al. 2018), ZDOCK (Pierce et al. 2014), and HADDOCK (van Zundert et al. 2015) were used to examine which structural information contributes to the accuracy of the model building of an antibody–antigen complex, such as the structural information of CDR loops, paratope (antigen-binding residues on an antibody CDR), antigen surface, and epitope (antibody-binding residues on an antigen) (Ambrosetti et al. 2020). The findings showed for all docking methods that the overall performance decreased without epitope information. They were improved by consideration of the low-resolution epitope information. Accurate modeling of the structure of the long CDR-H3 loop remains challenging. However, the flexible refinement of HADDOCK led to the improvement of the prediction accuracy of CDR-H3 loop conformations if the epitope information was available, even though the resolution of the information is low (Ambrosetti et al. 2020).

The AI-based prediction of antibody–antigen interfaces has been developed as PECAN (Pittala and Bailey-Kellogg 2020). Here, antibody and antigen structures are presented as the respective graphs. They are input to the neural network that consists of graph convolution, attention, and fully connected layers. The network discriminates antibody and antigen residues between interface and non-interface residues. The predictive accuracies of PECAN achieved when using the datasets provided with epitope (Krawczyk et al. 2014), and paratope prediction methods (Daberdaku and Ferrari 2019) were found to have higher precision and recall rates than these providers' methods. It is noteworthy that the providers' methods also predict epitope and paratope regions with high accuracies. Epitope prediction, using the program EpiPred, predicts epitopes based on geometric fitting and knowledge-based asymmetric antibody–antigen scoring (Krawczyk et al. 2014). The paratope prediction method uses a SVM classifier to distinguish interface surface patches from non-interface ones based on 3D Zernike descriptors that represent global and local protein surface shapes and physicochemical properties on the surfaces (Daberdaku and Ferrari 2019). Again, note that even recent AI-based approaches, such as AF2-Multimer, suffer from an inability to predict accurate antibody-antigen complex structures (as mentioned above).

To obtain information about antibody-specific epitopes, some improvements in the accuracy of docking and affinity predictions must be achieved. For this purpose, a large and non-redundant benchmark set for antibody–antigen docking and affinity prediction was constructed. It includes camelid nanobodies, therapeutic monoclonal antibodies, and broadly neutralizing antibodies that target viral glycoproteins (Guest et al. 2021).

Another possible approach is the search of similar regions on proteins to known antibody-binding epitopes. This approach is based on the fact that some antibodies can cross-reactively recognize different antigen proteins having similar surface regions in the structures and properties (Vieths et al. 2006; Negi and Braun 2017). Information about cross-reactivity might provide information about the repurposing of antibody drugs. We are striving to develop a database of known and putative epitopes on proteins (PoSSuMAg 2022), which has a similar scheme to that of the PoSSuM database (PoSSuM 2021), which is helpful to detect putative pockets that are similar to known ligand-binding sites on protein structures (Tabei et al. 2010; Ito et al. 2012b, a, 2015). Our new database presents information about putative epitopes that are similar to known epitopes on the antigen proteins in complex with antibodies, including antibody drugs. Information about known and putative epitopes on SARS-CoV-2 proteins will also be available. Although the current version includes information of putative epitopes on non-redundant protein structures, we expect to increase the data through future studies.

Concluding remarks

Here we have introduced and discussed a diverse range of computational methods, from docking-based to AI-based approaches, for PPI prediction. Along with the increase in

the amount of information related to protein sequences and structures, AI-based approaches, especially those based on evolutionary information and templates, are expected to become more powerful and useful. Yet, prediction of hetero PPIs leaves room for improvement. Particularly, antibody–antigen interaction predictions remain very limited in terms of their accuracy, although many sophisticated prediction methods have been developed (Ambrosetti et al. 2020).

Acknowledgements We thank the organizers of the issue for extending this opportunity to us.

Author contribution YT, YY, and KT contributed to the design of the study and wrote the manuscript. PPI predictions in this study were performed by YY. All authors read and approved the final manuscript.

Funding The study was supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP22ama121028.

Data Availability All the data associated with Fig. 1 are available from our figshare repository (Yamamori et al. 2022).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahdritz G, Bouatta N, Kadyan S, Xia Q, Gerecke W, O'Donnell TJ, Berenberg D, Fisk I, Zanichelli N, Zhang B, Nowaczynski A, Wang B, Stepniewska-Dziubinska MM, Zhang S, Ojewole A, Guney ME, Biderman S, Watkins AM, Ra S, Lorenzo PR, et al (2022) OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv 2022.11.20.517210. https://doi.org/10.1101/2022.11.20.517210

Ambrosetti F, Jiménez-García B, Roel-Touris J, Bonvin AMJJ (2020) Modeling antibody-antigen complexes by information-driven docking. Structure 28:119-129.e2. https://doi.org/10.1016/j.str.2019.10.011

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754

Basu S, Wallner B (2016) DockQ: a quality measure for protein-protein docking models. PLoS ONE 11:e0161879. https://doi.org/10.1371/journal.pone.0161879

Brenke R, Hall DR, Chuang G-Y, Comeau SR, Bohnuud T, Beglov D, Schueler-Furman O, Vajda S, Kozakov D (2012) Application of asymmetric statistical potentials to antibody-protein docking. Bioinformatics 28:2608–2614. https://doi.org/10.1093/bioinformatics/bts493

Bryant P, Pozzati G, Elofsson A (2022a) Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13:1265. https://doi.org/10.1038/s41467-022-28865-w

Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A (2022b) Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. BioRxiv. https://doi.org/10.1101/2022.03.12.484089

Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A, Scheltema RA, Bruce JE, Leitner A, Kundrotas P, Beltrao P, Elofsson A (2022) Towards a structurally resolved human protein interaction network. bioRxiv. https://doi.org/10.1101/2021.11.08.467664

CAPRI (2022) CAPRI: Critical Assessment of PRediction of Interactions. https://www.ebi.ac.uk/pdbe/complex-pred/capri/ (Accessed 2nd December 2022)

CAPRI Round 54 (2022) CASP15-CAPRI assembly prediction experiment. https://www.ebi.ac.uk/pdbe/complex-pred/capri/round/54/ (Accessed 2nd December 2022)

Daberdaku S, Ferrari C (2019) Antibody interface prediction with 3D Zernike descriptors and SVM. Bioinformatics 35:1870–1876. https://doi.org/10.1093/bioinformatics/bty918

Das S, Chakrabarti S (2021) Classification and prediction of protein-protein interaction interface using machine learning algorithm. Sci Rep 11:1761. https://doi.org/10.1038/s41598-020-80900-2

DeLano WL (2002) Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12:14–20. https://doi.org/10.1016/S0959-440X(02)00283-X

Desta IT, Porter KA, Xia B, Kozakov D, Vajda S (2020) Performance and its limits in rigid body protein-protein docking. Structure 28:1071-1081.e3. https://doi.org/10.1016/j.str.2020.06.006

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P et al (2021) Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034

Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, Zhou H-X, Mitchell JC, Demerdash ONA, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko J, Lee H, Seok C, Bourquard T, Bernauer J et al (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414:289–302. https://doi.org/10.1016/j.jmb.2011.09.031

Fukasawa Y, Tomii K (2019) Accurate classification of biological and non-biological interfaces in protein crystal structures using subtle covariation signals. Sci Rep 9:12603. https://doi.org/10.1038/s41598-019-48913-8

Gao M, Nakajima An D, Parks JM, Skolnick J (2022) AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. Nat Commun 13:1744. https://doi.org/10.1038/s41467-022-29394-2

Geng C, Jung Y, Renaud N, Honavar V, Bonvin AMJJ, Xue LC (2020) iScore: a novel graph kernel-based function for scoring protein–protein docking models. Bioinformatics 36:112–121. https://doi.org/10.1093/bioinformatics/btz496

Ghani U, Desta I, Jindal A, Khan O, Jones G, Kotelnikov S, Padhorny D, Vajda S, Kozakov D (2021) Improved docking of protein

models by a combination of Alphafold2 and ClusPro. bioRxiv. https://doi.org/10.1101/2021.09.07.459290
Grosdidier S, Fernández-Recio J (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics 9:447. https://doi.org/10.1186/1471-2105-9-447
Guest JD, Vreven T, Zhou J, Moal I, Jeliazkov JR, Gray JJ, Weng Z, Pierce BG (2021) An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. Structure 29:606-621.e5. https://doi.org/10.1016/J.STR.2021.01.005
Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, Stancheva VG, Li X-H, Liu K, Zheng Z, Barrero DJ, Roy U, Kuper J, Fernández IS, Szakal B, Branzei D et al (2021) Computed structures of core eukaryotic protein complexes. Science 374:eabm4805. https://doi.org/10.1126/science.abm4805
Ito J-I, Tabei Y, Shimizu K, Tomii K, Tsuda K (2012a) PDB-scale analysis of known and putative ligand-binding sites with structural sketches. Proteins Struct Funct Bioinforma 80:747–763. https://doi.org/10.1002/prot.23232
Ito J-I, Tabei Y, Shimizu K, Tsuda K, Tomii K (2012b) PoSSuM: a database of similar protein-ligand binding and putative pockets. Nucleic Acids Res 40:D541–D548. https://doi.org/10.1093/nar/gkr1130
Ito J, Ikeda K, Yamada K, Mizuguchi K, Tomii K (2015) PoSSuM vol 2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. Nucleic Acids Res 43:D392–D398. https://doi.org/10.1093/nar/gku1144
Jiménez-García B, Pons C, Svergun DI, Bernadó P, Fernández-Recio J (2015) pyDockSAXS: protein–protein complex structure by SAXS and computational docking. Nucleic Acids Res 43:W356–W361. https://doi.org/10.1093/nar/gkv368
Jiménez-García B, Roel-Touris J, Romero-Durana M, Vidal M, Jiménez-González D, Fernández-Recio J (2018) LightDock: a new multi-scale approach to protein–protein docking. Bioinformatics 34:49–55. https://doi.org/10.1093/bioinformatics/btx555
Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63:31–65. https://doi.org/10.1016/0079-6107(94)00008-w
Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93:13–20. https://doi.org/10.1073/pnas.93.1.13
Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J Mol Biol 362:861–875. https://doi.org/10.1016/j.jmb.2006.07.072
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
Kanamori E, Murakami Y, Tsuchiya Y, Standley DM, Nakamura H, Kinoshita K (2007) Docking of protein molecular surfaces with evolutionary trace analysis. Proteins Struct Funct Bioinforma 69:832–838. https://doi.org/10.1002/prot.21737
Ko J, Lee J (2021) Can AlphaFold2 predict protein-peptide complex structures accurately? bioRxiv. https://doi.org/10.1101/2021.07.27.453972
Krawczyk K, Liu X, Baker T, Shi J, Deane CM (2014) Improving B-cell epitope prediction and its application to global antibody-antigen docking. Bioinformatics 30:2288–2294. https://doi.org/10.1093/bioinformatics/btu190
Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, Schwede T, Capitani G (2018) Assessment of protein assembly prediction in CASP12. Proteins Struct Funct Bioinforma 86:247–256. https://doi.org/10.1002/prot.25408
Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, Zhao D, Zeng J (2021) A deep-learning framework for multi-level peptide-protein interaction prediction. Nat Commun 12:5465. https://doi.org/10.1038/s41467-021-25772-4
Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084. https://doi.org/10.1002/prot.22818
Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond ASJ, Karaca E, Schmitz C, van Dijk M, Bonvin AMJJ, Eisenstein M, Jiménez-García B, Grosdidier S, Solernou A, Pérez-Cano L, Pallara C, Fernández-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ et al (2014) Blind prediction of interfacial water positions in CAPRI. Proteins Struct Funct Bioinforma 82:620–632. https://doi.org/10.1002/prot.24439
Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Lee GR, Seok C, Qin S et al (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins Struct Funct Bioinforma 84:323–348. https://doi.org/10.1002/prot.25007
Lensink MF, Velankar S, Baek M, Heo L, Seok C, Wodak SJ (2018) The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. Proteins Struct Funct Bioinforma 86:257–273. https://doi.org/10.1002/prot.25419
Lensink MF, Nadzirin N, Velankar S, Wodak SJ (2020) Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. Proteins Struct Funct Bioinforma 88:916–938. https://doi.org/10.1002/prot.25870
Li Z, Liu X, Chen W, Shen F, Bi H, Ke G, Zhang L, Technology DP (2022a) Uni-Fold: an open-source platform for developing protein folding models beyond AlphaFold. bioRxiv. https://doi.org/10.1101/2022.08.04.502811
Li Z, Yang S, Liu X, Chen W, Wen H, Shen F, Ke G, Zhang L, Technology DP (2022b) Uni-Fold Symmetry: harnessing symmetry in folding large protein complexes. bioRxiv. https://doi.org/10.1101/2022.08.30.505833
Lim H, Chun J, Jin X, Kim J, Yoon J, No KT (2019) Investigation of protein-protein interactions and hot spot region between PD-1 and PD-L1 by fragment molecular orbital method. Sci Rep 9:16727. https://doi.org/10.1038/s41598-019-53216-z
Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Dos A, Costa S, Fazel-Zarandi M, Sercu T, Candido S, Rives A, Ai M (2022) Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. https://doi.org/10.1101/2022.07.20.500902
Liu S, Zhang J, Chu H, Wang M, Xue B, Ni N, Yu J, Xie Y, Chen Z, Chen M, Liu Y, Patra P, Xu F, Chen J, Wang Z, Yang L, Yu F, Chen L, Gao YQ (2022) PSP: million-level protein sequence dataset for protein structure prediction. ArXiv. https://doi.org/10.48550/arxiv.2206.12240
Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, Shi J (2020) Recent advances in the development of protein-protein interactions modulators: mechanisms and clinical trials. Signal Transduct Target Ther 5:213. https://doi.org/10.1038/s41392-020-00315-3
Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold: making protein folding accessible to all. Nat Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1
Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JPGLM, Trellet M, Bonvin AMJJ, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L,

Pallara C, Jimenez B, Fernandez-Recio J, Flores S et al (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. Proteins Struct Funct Bioinforma 81:1980–1987. https://doi.org/10.1002/prot.24356

Moriwaki Y (2021) Twitter post: AlphaFold2 can also predict hetero-complexes. all you have to do is input the two sequences you want to predict and connect them with a long linker. In: https://twitter.com/Ag_smith/status/1417063635000598528

Nakamura T, Oda T, Fukasawa Y, Tomii K (2018) Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments. Proteins Struct Funct Bioinforma. https://doi.org/10.1002/prot.25432

Negi SS, Braun W (2017) Cross-React: a new structural bioinformatics method for predicting allergen cross-reactivity. Bioinformatics 33:1014–1020. https://doi.org/10.1093/bioinformatics/btw767

Ovek D, Abali Z, Zeylan ME, Keskin O, Gursoy A, Tuncbag N (2022) Artificial intelligence based methods for hot spot prediction. Curr Opin Struct Biol 72:209–218. https://doi.org/10.1016/j.sbi.2021.11.003

Ozdemir ES, Gursoy A, Keskin O (2018) Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. Bioinformatics 34:i795–i801. https://doi.org/10.1093/bioinformatics/bty569

PIACO 2019 Protein Interface Analysis using COvarying signals. https://github.com/yfukasawa/piaco (Accessed 2nd December 2022)

Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 30:1771–1773. https://doi.org/10.1093/bioinformatics/btu097

Pittala S, Bailey-Kellogg C (2020) Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics 36:3996–4003. https://doi.org/10.1093/bioinformatics/btaa263

PoSSuM 2021 Pocket Similarity Search using Multiple-Sketches https://possum.cbrc.pj.aist.go.jp/PoSSuM/ (Accessed 2nd December 2022)

PoSSuMAg (2022) Pocket Similarity Search using Multiple-Sketches (Antigen) (in preparation)

Poupon A, Janin J (2010) Analysis and prediction of protein quaternary structure. Methods Mol Biol 609:349–364. https://doi.org/10.1007/978-1-60327-241-4_20

Rosell M, Fernández-Recio J (2020) Docking-based identification of small-molecule binding sites at protein-protein interfaces. Comput Struct Biotechnol J 18:3750–3761. https://doi.org/10.1016/j.csbj.2020.11.029

Ruffolo JA, Chu L-S, Mahajan SP, Gray JJ (2022) Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. bioRxiv. https://doi.org/10.1101/2022.04.20.488972

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33:W363-367. https://doi.org/10.1093/nar/gki481

Smith GR, Sternberg MJE (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12:28–35. https://doi.org/10.1016/S0959-440x(02)00285-3

Tabei Y, Uno T, Sugiyama M, Tsuda K (2010) Single versus multiple sorting for all pairs similarity search. In: The Second Asian Conference on Machine Learning (ACML2010), Tokyo, Japan. pp 145–160

Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17:284–285. https://doi.org/10.1093/bioinformatics/17.3.284

Tomii K, Akiyama Y (2004) FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 20:594–595. https://doi.org/10.1093/bioinformatics/btg474

Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O (2022) Harnessing protein folding neural networks for peptide-protein docking. Nat Commun 13:176. https://doi.org/10.1038/s41467-021-27838-9

Tsuchiya Y, Kinoshita K, Ito N, Nakamura H (2006a) PreBI: prediction of biological interfaces of proteins in crystals. Nucleic Acids Res 34:W320-324. https://doi.org/10.1093/nar/gkl267

Tsuchiya Y, Kinoshita K, Nakamura H (2006b) Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. Protein Eng Des Sel 19:421–429. https://doi.org/10.1093/protein/gzl026

van Noort CW, Honorato RV, Bonvin AMJJ (2021) Information-driven modeling of biomolecular complexes. Curr Opin Struct Biol 70:70–77. https://doi.org/10.1016/j.sbi.2021.05.003

van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, van Dijk M, de Vries SJ, Bonvin AMJJ (2015) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol 428:720–725. https://doi.org/10.1016/j.jmb.2015.09.014

Vieths S, Scheurer S, Ballmer-Weber B (2006) Current understanding of cross-reactivity of food allergens and pollen. Ann N Y Acad Sci 964:47–68. https://doi.org/10.1111/j.1749-6632.2002.tb04132.x

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, Bonvin AMJJ, Weng Z (2015) Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427:3031–3041. https://doi.org/10.1016/J.JMB.2015.07.016

Wang X, Terashi G, Christoffer CW, Zhu M, Kihara D (2020) Protein docking model evaluation by 3D deep convolutional neural networks. Bioinformatics 36:2113–2118. https://doi.org/10.1093/bioinformatics/btz870

Wang G, Fang X, Wu Z, Liu Y, Xue Y, Xiang Y, Yu D, Wang F, Ma Y (2022) HelixFold: an efficient implementation of AlphaFold2 using PaddlePaddle. ArXiv. https://doi.org/10.48550/arxiv.2207.05477

Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B, Ma J, Peng J (2022) High-resolution de novo structure prediction from primary sequence. bioRxiv. https://doi.org/10.1101/2022.07.21.500999

Yamamori Y, Tsuchiya Y, Tomii K (2022) PPI prediction results for six CASP14 targets using AF2-related methods. https://doi.org/10.6084/m9.figshare.21716330

Yin R, Feng BY, Varshney A, Pierce BG, Brian Pierce CG (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Sci 31:e4379. https://doi.org/10.1002/PRO.4379

Yu J, Vavrusa M, Andreani J, Rey J, Tufféry P, Guerois R (2016) InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. Nucleic Acids Res 44:W542–W549. https://doi.org/10.1093/nar/gkw340

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

D.12 Machine learning

Further explanations for Chapter 12.

D.12.1 Basics

- Chicco (2017) — Practical advice on the whole workflow, from data preparation and splitting to hyperparameter tuning, imbalanced classes and the choice of evaluation metric. A short and accessible complement to the section on generalisation, overfitting and evaluation. *(tutorial; CC BY 4.0)*

  Ten quick tips for machine learning in computational biology. Chicco (2017). *BioData Mining*. doi:10.1186/s13040-017-0155-3

  > Machine learning has become a pivotal tool for many projects in computational biology, bioinformatics, and health informatics. Nevertheless, beginners and biomedical researchers often do not have enough experience to run a data mining project effectively, and therefore can follow incorrect practices, that may lead to common mistakes or over-optimistic results. With this review, we present ten quick tips to take advantage of machine learning in any computational biology context, by avoiding some common errors that we observed hundreds of times in multiple bioinformatics projects. We believe our ten suggestions can strongly help any machine learning practitioner to carry on a successful project in computational biology and related sciences.

- Miller et al. (2024) — Overviews evaluation metrics for clustering, classification and regression, their advantages and disadvantages, and common pitfalls in model evaluation, with examples from genomics. Useful for the evaluation subsection. *(review; CC BY 4.0)*

  A review of model evaluation metrics for machine learning in genetics and genomics. Miller et al. (2024). *Frontiers in Bioinformatics*. doi:10.3389/fbinf.2024.1457619

  > Machine learning (ML) has shown great promise in genetics and genomics where large and complex datasets have the potential to provide insight into many aspects of disease risk, pathogenesis of genetic disorders, and prediction of health and wellbeing. However, with this possibility there is a responsibility to exercise caution against biases and inflation of results that can have harmful unintended impacts. Therefore, researchers must understand the metrics used to evaluate ML models which can influence the critical interpretation of results. In this review we provide an overview of ML metrics for clustering, classification, and regression and highlight the advantages and disadvantages of each. We also detail common pitfalls that occur during model evaluation. Finally, we provide examples of how researchers can assess and utilise the results of ML models, specifically from a genomics perspective.

D.12.2 A short history of machine learning in bioinformatics

- Ben-Hur et al. (2008) — A tutorial introduction to support vector machines and kernels, including kernels for biological sequences, written for computational biologists. It explains the margin, soft margins and model selection clearly. *(tutorial; CC BY 4.0)*

  Support Vector Machines and Kernels for Computational Biology. Ben-Hur et al. (2008). *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1000173

- Yoon (2009) — Introduces hidden Markov models, the Viterbi, forward and Baum–Welch algorithms, and profile HMMs, with applications in sequence analysis. Gives more detail than the section on how HMMs are decoded and trained. *(review; CC BY (Europe PMC))*

  Hidden Markov Models and their Applications in Biological Sequence Analysis. Yoon (2009). *Current Genomics*. doi:10.2174/138920209789177575

> Hidden Markov models (HMMs) have been extensively used in biological sequence analysis. In this paper, we give a tutorial review of HMMs and their applications in a variety of problems in molecular biology. We especially focus on three types of HMMs: the profile-HMMs, pair-HMMs, and context-sensitive HMMs. We show how these HMMs can be used to solve various sequence analysis problems, such as pairwise and multiple sequence alignments, gene annotation, classification, similarity search, and many others.

- Touw et al. (2012) — Describes how random forests work and how they are used in the life sciences, including variable importance measures and their pitfalls. Licence is non-commercial, so it is suitable only as a link. *(review; CC BY-NC (link only))*

  Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. Touw et al. (2012). *Briefings in Bioinformatics*. doi:10.1093/bib/bbs034

  > In the Life Sciences 'omics' data is increasingly generated by different high-throughput technologies. Often only the integration of these data allows uncovering biological insights that can be experimentally validated or mechanistically modelled, i.e. sophisticated computational approaches are required to extract the complex non-linear trends present in omics data. Classification techniques allow training a model based on variables (e.g. SNPs in genetic association studies) to separate different classes (e.g. healthy subjects versus patients). Random Forest (RF) is a versatile classification algorithm suited for the analysis of these large data sets. In the Life Sciences, RF is popular because RF classification models have a high-prediction accuracy and provide information on importance of variables for classification. For omics data, variables or conditional relations between variables are typically important for a subset of samples of the same class. For example: within a class of cancer patients certain SNP combinations may be important for a subset of patients that have a specific subtype of cancer, but not important for a different subset of patients. These conditional relationships can in principle be uncovered from the data with RF as these are implicitly taken into account by the algorithm during the creation of the classification model. This review details some of the to the best of our knowledge rarely or never used RF properties that allow maximizing the biological insights that can be extracted from complex omics data sets using RF.

D.12.3 Deep learning

- Angermueller et al. (2016) — An introduction to deep neural networks for biologists, covering training, regularisation, convolutional networks and applications in regulatory genomics and imaging, with practical recommendations. Predates transformers and language models. *(review; CC BY 4.0)*

  Deep learning for computational biology. Angermueller et al. (2016). *Molecular Systems Biology*. doi:10.15252/msb.20156651

  > Technological advances in genomics and imaging have led to an explosion of molecular and cellular profiling data from large numbers of samples. This rapid increase in biological data dimension and acquisition rate is challenging conventional analysis strategies. Modern machine learning methods, such as deep learning, promise to leverage very large data sets for finding hidden structure within them, and for making accurate predictions. In this review, we discuss applications of this new breed of analysis approaches in regulatory genomics and cellular imaging. We provide background of what deep learning is, and the settings in which it can be successfully applied to derive biological insights. In addition to presenting specific applications and providing tips for practical use, we also highlight possible pitfalls and limitations to guide computational biologists when and how to make the most use of this new technology.

- Lee et al. (2022) — Short, practical guidance on when deep learning is appropriate, comparing with simple baselines, avoiding overfitting and leakage, and interpreting and sharing models. Complements the subsections on training and on interpretability and reporting. *(tutorial; CC BY 4.0)*

  Ten quick tips for deep learning in biology. Lee et al. (2022). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1009803

- Zhang et al. (2023) — Explains the transformer architecture and self-supervised pre-training, and surveys their use for protein, DNA and RNA sequences and other biological data. Relevant to the subsection on language models for proteins and DNA. *(review; CC BY 4.0)*

  Applications of transformer-based language models in bioinformatics: a survey. Zhang et al. (2023). *Bioinformatics Advances*. doi:10.1093/bioadv/vbad001

  > The transformer-based language models, including vanilla transformer, BERT and GPT-3, have achieved revolutionary breakthroughs in the field of natural language processing (NLP). Since there are inherent similarities between various biological sequences and natural languages, the remarkable interpretability and adaptability of these models have prompted a new wave of their application in bioinformatics research. To provide a timely and comprehensive review, we introduce key developments of transformer-based language models by describing the detailed structure of transformers and summarize their contribution to a wide range of bioinformatics research from basic sequence analysis to drug discovery. While transformer-based applications in bioinformatics are diverse and multifaceted, we identify and discuss the common challenges, including heterogeneity of training data, computational expense and model interpretability, and opportunities in the context of bioinformatics research. We hope that the broader community of NLP researchers, bioinformaticians and biologists will be brought together to foster future research and development in transformer-based language models, and inspire novel bioinformatics applications that are unattainable by traditional methods. Supplementary information Supplementary data are available at Bioinformatics Advances online.

D.12.4 Generative models

- Lopez et al. (2020) — Explains the principles of deep generative models, in particular variational autoencoders and GANs, and how they are used for single-cell data, sequences and images. Predates the use of diffusion models in biology. *(review; CC BY 4.0)*

  Enhancing scientific discoveries in molecular biology with deep generative models. Lopez et al. (2020). *Molecular Systems Biology*. doi:10.15252/msb.20199198

  > Generative models provide a well-established statistical framework for evaluating uncertainty and deriving conclusions from large data sets especially in the presence of noise, sparsity, and bias. Initially developed for computer vision and natural language processing, these models have been shown to effectively summarize the complexity that underlies many types of data and enable a range of applications including supervised learning tasks, such as assigning labels to images; unsupervised learning tasks, such as dimensionality reduction; and out-of-sample generation, such as de novo image synthesis. With this early success, the power of generative models is now being increasingly leveraged in molecular biology, with applications ranging from designing new molecules with properties of interest to identifying deleterious mutations in our genomes and to dissecting transcriptional variability between single cells. In this review, we provide a brief overview of the technical notions behind generative models and their implementation with deep learning techniques. We then describe several different ways in which these models can be utilized in practice, using several recent applications in molecular biology as examples.

- Strokach and Kim (2022) — A concise review of deep generative models of protein sequences and structures and how they are used, together with predictive models, to propose and select new protein designs. *(review; CC BY 4.0)*

  Deep generative modeling for protein design. Strokach & Kim (2022). *Current Opinion in Structural Biology*. doi:10.1016/j.sbi.2021.11.008

  > Deep learning approaches have produced substantial breakthroughs in fields such as image classification and natural language processing and are making rapid inroads in the area of protein design. Many generative models of proteins have been developed that encompass all known protein sequences, model specific protein families, or extrapolate the dynamics of individual proteins. Those generative models can learn protein representations that are often more informative of protein structure and function than hand-engineered features. Furthermore, they can be used to quickly propose millions of novel proteins that resemble the native counterparts in terms of expression level, stability, or other attributes. The protein design process can further be guided by discriminative oracles to select candidates with the highest probability of having the desired properties. In this review, we discuss five classes of generative

models that have been most successful at modeling proteins and provide a framework for model guided protein design.

- Mardikoraem et al. (2023) — Describes the architecture and mathematics of self-supervised and generative models for protein sequences and their use in protein engineering and fitness prediction. *(review; CC BY 4.0)*

  Generative models for protein sequence modeling: recent advances and future directions. Mardikoraem et al. (2023). *Briefings in Bioinformatics*. doi:10.1093/bib/bbad358

  > The widespread adoption of high-throughput omics technologies has exponentially increased the amount of protein sequence data involved in many salient disease pathways and their respective therapeutics and diagnostics. Despite the availability of large-scale sequence data, the lack of experimental fitness annotations underpins the need for self-supervised and unsupervised machine learning (ML) methods. These techniques leverage the meaningful features encoded in abundant unlabeled sequences to accomplish complex protein engineering tasks. Proficiency in the rapidly evolving fields of protein engineering and generative AI is required to realize the full potential of ML models as a tool for protein fitness landscape navigation. Here, we support this work by (i) providing an overview of the architecture and mathematical details of the most successful ML models applicable to sequence data (e.g. variational autoencoders, autoregressive models, generative adversarial neural networks, and diffusion models), (ii) guiding how to effectively implement these models on protein sequence data to predict fitness or generate high-fitness sequences and (iii) highlighting several successful studies that implement these techniques in protein engineering (from paratope regions and subcellular localization prediction to high-fitness sequences and protein design rules generation). By providing a comprehensive survey of model details, novel architecture developments, comparisons of model applications, and current challenges, this study intends to provide structured guidance and robust framework for delivering a prospective outlook in the ML-driven protein engineering field.

D.12.5 Foundation models

- Li et al. (2024) — Reviews foundation models for sequences, structures and omics data, how they are pre-trained and adapted to downstream tasks, and their current limitations. *(review; CC BY 4.0)*

  Progress and opportunities of foundation models in bioinformatics. Li et al. (2024). *Briefings in Bioinformatics*. doi:10.1093/bib/bbae548

  > Bioinformatics has undergone a paradigm shift in artificial intelligence (AI), particularly through foundation models (FMs), which address longstanding challenges in bioinformatics such as limited annotated data and data noise. These AI techniques have demonstrated remarkable efficacy across various downstream validation tasks, effectively representing diverse biological entities and heralding a new era in computational biology. The primary goal of this survey is to conduct a general investigation and summary of FMs in bioinformatics, tracing their evolutionary trajectory, current research landscape, and methodological frameworks. Our primary focus is on elucidating the application of FMs to specific biological problems, offering insights to guide the research community in choosing appropriate FMs for tasks like sequence analysis, structure prediction, and function annotation. Each section delves into the intricacies of the targeted challenges, contrasting the architectures and advancements of FMs with conventional methods and showcasing their utility across different biological domains. Further, this review scrutinizes the hurdles and constraints encountered by FMs in biology, including issues of data noise, model interpretability, and potential biases. This analysis provides a theoretical groundwork for understanding the circumstances under which certain FMs may exhibit suboptimal performance. Lastly, we outline prospective pathways and methodologies for the future development of FMs in biological research, facilitating ongoing innovation in the field. This comprehensive examination not only serves as an academic reference but also as a roadmap for forthcoming explorations and applications of FMs in biology.

- Guo et al. (2025) — A broad overview of language and foundation models for proteins, nucleic acids and single-cell data, with discussion of evaluation and open problems. *(review; CC BY 4.0)*

  Foundation models in bioinformatics. Guo et al. (2025). *National Science Review*. doi:10.1093/nsr/nwaf028

> With the adoption of foundation models (FMs), artificial intelligence (AI) has become increasingly significant in bioinformatics and has successfully addressed many historical challenges, such as pre-training frameworks, model evaluation and interpretability. FMs demonstrate notable proficiency in managing large-scale, unlabeled datasets, because experimental procedures are costly and labor intensive. In various downstream tasks, FMs have consistently achieved noteworthy results, demonstrating high levels of accuracy in representing biological entities. A new era in computational biology has been ushered in by the application of FMs, focusing on both general and specific biological issues. In this review, we introduce recent advancements in bioinformatics FMs employed in a variety of downstream tasks, including genomics, transcriptomics, proteomics, drug discovery and single-cell analysis. Our aim is to assist scientists in selecting appropriate FMs in bioinformatics, according to four model types: language FMs, vision FMs, graph FMs and multimodal FMs. In addition to understanding molecular landscapes, AI technology can establish the theoretical and practical foundation for continued innovation in molecular biology.

- Baek et al. (2025) — Explains how single-cell foundation models, typically transformers pre-trained on large collections of transcriptomes, are built and used, and discusses their current challenges. *(review; CC BY 4.0)*

  Single-cell foundation models: bringing artificial intelligence into cell biology. Baek et al. (2025). *Experimental & Molecular Medicine*. doi:10.1038/s12276-025-01547-5

  > A foundation model, a large-scale deep learning model pretrained on vast datasets, has revolutionized data interpretation through self-supervised learning with capacity for various downstream tasks. Concurrently, single-cell genomics is in urgent need of unified frameworks capable of integrating and comprehensively analyzing the rapidly expanding data repositories. Inspired by advances in foundation models, researchers have extended these techniques to single-cell analysis, giving rise to single-cell foundation models (scFMs). Typically, these models use transformer architectures to incorporate diverse omics data and extract latent patterns at both cell and gene/feature levels for the analysis of cellular heterogeneity and complex regulatory networks. Despite their promise, scFMs face challenges including the nonsequential nature of omics data, inconsistency in data quality and the computational intensity required for training and fine-tuning. Furthermore, interpreting the biological relevance of latent embeddings and model representations remains nontrivial. Here we provide an overview of scFMs, highlighting their key concepts and applications across downstream tasks. We critically assess the current limitations and propose future directions aimed at enhancing the robustness, interpretability and scalability of scFMs. Ultimately, addressing these challenges will be crucial for establishing scFMs as pivotal tools in advancing single-cell genomics and unlocking deeper insights into cellular function and disease mechanisms.

D.12.6 AI in research practice

- Smith et al. (2024) — Practical rules for using LLMs responsibly in research, including verifying output, protecting data, disclosure and reproducibility. Matches the risks listed in the section. *(tutorial; CC BY 4.0)*

  Ten simple rules for using large language models in science, version 1.0. Smith et al. (2024). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1011767

  > Box 1-An LLM suggests 10 simple rules governing its own usage in scienceHere, we ask ChatGPT to generate a list of 10 simple rules for using LLMs in science in order to illustrate its ability to provide detailed, human-like responses to short prompts.Q1: Can you provide me with a list of ten simple rules for using large language models in science?R1 (ChatGPT 3.5): Certainly!Here are ten simple rules for using large language models in science:1. Understand the limitations: Recognize that large language models like GPT-3 are powerful but not infallible.They may produce accurate-sounding information, but they can also generate false or biased content.Exercise caution and critical thinking.2. Validate and verify: Always cross-reference information generated by language models with reliable sources.Verify the accuracy of claims, check for

- Lubiana et al. (2023) — Advice on using chatbots for coding and other tasks in computational biology, with emphasis on checking generated code and text. *(tutorial; CC BY 4.0)*

  Ten quick tips for harnessing the power of ChatGPT in computational biology. Lubiana et al. (2023). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1011319

> The rise of advanced chatbots, such as ChatGPT, has stirred excitement and curiosity in the scientific community.Powered by large language models (LLMs) based on generative pretrained transformers (GPTs)-specifically GPT-3.5 and GPT-4-ChatGPT is considered a general-purpose technology with the potential to impact the job market and research endeavors in numerous fields [1].Although similar models have been fine-tuned for biology-specific projects, including text-based analysis and biological sequence decoding [2,3], ChatGPT provides a natural interface for bioinformaticians to begin using LLMs in their activities.This tool is already accelerating various activities undertaken by computational biologists, ranging from data cleaning to interpreting results and publishing.However, with great power comes great responsibility.As scientists, we must harness the full potential of ChatGPT while adhering to ethical guidelines and avoiding pitfalls associated with the technology.Here, we provide 10 insightful tips designed to help computational biologists optimize their workflows with ChatGPT, ranging from basic prompts to more advanced techniques.Although our primary focus is on the current ChatGPT/GPT-4 model, we believe that these tips will remain relevant for future iterations of the technology, as well as other LLMs and chatbots (such as Meta's LLaMa and Google's Bard) [4,5].We invite you to explore our 10 tips (summarized in Fig 1) aimed at effectively utilizing ChatGPT to advance computational biology research while maintaining a strong commitment to research integrity. Tip 1: Embrace the technology and be ready for noveltyChatGPT, a powerful tool for coding and academic writing tasks, is rapidly gaining traction in the scientific community.While exercising critical judgment and not blindly accepting everything it produces is important, incorporating ChatGPT into your workflow can undoubtedly

- Helmy et al. (2025) — Recent guidance on using generative AI tools in scientific work, covering verification, transparency and ethical issues. *(tutorial; CC BY 4.0)*

  Ten simple rules for optimal and careful use of generative AI in science. Helmy et al. (2025). *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1013588

  > Modern AI technologies leverage natural language processing (NLP), a subfield of AI dedicated to understanding, interpreting, and generating human language for developing large language models (LLMs), which have significantly advanced the capabilities of AI systems. These models can perform complex language tasks such as text generation, summarization, translation, and sentiment analysis, with unprecedented accuracy. The two main kinds of pre-training LLMs are the BERT-like models (e.g., BioBERT, proteinBERT, and PubMedBERT used primarily for language understanding; and the GPT-like models (e.g., BioGPT and ChatGPT-4o) used primarily for language generation

D.12.7 Integrating data

- Subramanian et al. (2020) — An overview of strategies and tools for integrating multi-omics data, organised by the question being asked (disease subtyping, biomarker prediction, insight into disease biology). *(review; CC BY-NC 4.0)*

  Multi-omics Data Integration, Interpretation, and Its Application. Subramanian et al. (2020). *Bioinformatics and Biology Insights*. doi:10.1177/1177932219899051

  > To study complex biological processes holistically, it is imperative to take an integrative approach that combines multi-omics data to highlight the interrelationships of the involved biomolecules and their functions. With the advent of high-throughput techniques and availability of multi-omics data generated from a large set of samples, several promising tools and methods have been developed for data integration and interpretation. In this review, we collected the tools and methods that adopt integrative approach to analyze multiple omics data and summarized their ability to address applications such as disease subtyping, biomarker prediction, and deriving insights into the data. We provide the methodology, use-cases, and limitations of these tools; brief account of multi-omics data repositories and visualization portals; and challenges associated with multi-omics data integration.

- Argelaguet et al. (2018) — The MOFA paper; a clear explanation of how latent factors shared across data types are learned and interpreted. *(primary study; CC BY 4.0)*

  Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Argelaguet et al. (2018). *Molecular Systems Biology*. doi:10.15252/msb.20178124

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and ex vivo drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

The following open-access paper(s) are reproduced under their original licences.

Review

OPEN ACCESS

molecular systems biology

Deep learning for computational biology

Christof Angermueller[1,†], Tanel Pärnamaa[2,3,†], Leopold Parts[2,3,*] & Oliver Stegle[1,**]

Abstract

Technological advances in genomics and imaging have led to an explosion of molecular and cellular profiling data from large numbers of samples. This rapid increase in biological data dimension and acquisition rate is challenging conventional analysis strategies. Modern machine learning methods, such as deep learning, promise to leverage very large data sets for finding hidden structure within them, and for making accurate predictions. In this review, we discuss applications of this new breed of analysis approaches in regulatory genomics and cellular imaging. We provide background of what deep learning is, and the settings in which it can be successfully applied to derive biological insights. In addition to presenting specific applications and providing tips for practical use, we also highlight possible pitfalls and limitations to guide computational biologists when and how to make the most use of this new technology.

Keywords cellular imaging; computational biology; deep learning; machine learning; regulatory genomics

DOI 10.15252/msb.20156651 | Received 11 April 2016 | Revised 2 June 2016 | Accepted 6 June 2016

Mol Syst Biol. (2016) 12: 878

Introduction

Machine learning methods are general-purpose approaches to learn functional relationships from data without the need to define them *a priori* (Hastie *et al*, 2005; Murphy, 2012; Michalski *et al*, 2013). In computational biology, their appeal is the ability to derive predictive models without a need for strong assumptions about underlying mechanisms, which are frequently unknown or insufficiently defined. As a case in point, the most accurate prediction of gene expression levels is currently made from a broad set of epigenetic features using sparse linear models (Karlic *et al*, 2010; Cheng *et al*, 2011) or random forests (Li *et al*, 2015); how the selected features determine the transcript levels remains an active research topic. Predictions in genomics (Libbrecht & Noble, 2015; Märtens *et al*, 2016), proteomics (Swan *et al*, 2013), metabolomics (Kell, 2005) or sensitivity to compounds (Eduati *et al*, 2015) all rely on machine learning approaches as a key ingredient.

Most of these applications can be described within the canonical machine learning workflow, which involves four steps: data cleaning and pre-processing, feature extraction, model fitting and evaluation (Fig 1A). It is customary to denote one data sample, including all covariates and features as *input* x (usually a vector of numbers), and label it with its response variable or *output* value y (usually a single number) when available.

A supervised machine learning model aims to learn a function $f(x) = y$ from a list of training pairs (x_1,y_1), (x_2,y_2), ... for which data are recorded (Fig 1B). One typical application in biology is to predict the viability of a cancer cell line when exposed to a chosen drug (Menden *et al*, 2013; Eduati *et al*, 2015). The input features (x) would capture somatic sequence variants of the cell line, chemical make-up of the drug and its concentration, which together with the measured viability (output label y) can be used to train a support vector machine, a random forest classifier or a related method (functional relationship f). Given a new cell line (unlabelled data sample x^*) in the future, the learnt function predicts its survival (output label y^*) by calculating $f(x^*)$, even if f resembles more of a black box, and its inner workings of why particular mutation combinations influence cell growth are not easily interpreted. Both regression (where y is a real number) and classification (where y is a categorical class label) can be viewed in this way. As a counterpart, unsupervised machine learning approaches aim to discover patterns from the data samples x themselves, without the need for output labels y. Methods such as clustering, principal component analysis and outlier detection are typical examples of unsupervised models applied to biological data.

The inputs x, calculated from the raw data, represent what the model "sees about the world", and their choice is highly problem-specific (Fig 1C). Deriving most informative features is essential for performance, but the process can be labour-intensive and requires domain knowledge. This bottleneck is especially limiting for high-dimensional data; even computational feature selection methods do not scale to assess the utility of the vast number of possible input combinations. A major recent advance in machine learning is automating this critical step by learning a suitable representation of the data with *deep artificial neural networks* (Bengio *et al*, 2013; LeCun *et al*, 2015; Schmidhuber, 2015) (Fig 1D). Briefly, a deep neural network takes the raw data at the lowest (input) layer and transforms them into increasingly abstract feature representations by successively combining outputs from the preceding layer in a data-driven manner, encapsulating highly complicated functions in the

1 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
2 Department of Computer Science, University of Tartu, Tartu, Estonia
3 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
*Corresponding author. Tel: +44 1223 834 244; E-mail: leopold.parts@sanger.ac.uk
**Corresponding author. Tel: +44 1223 494 101; E-mail: oliver.stegle@ebi.ac.uk
† These authors contributed equally to this work

© 2016 The Authors. Published under the terms of the CC BY 4.0 license

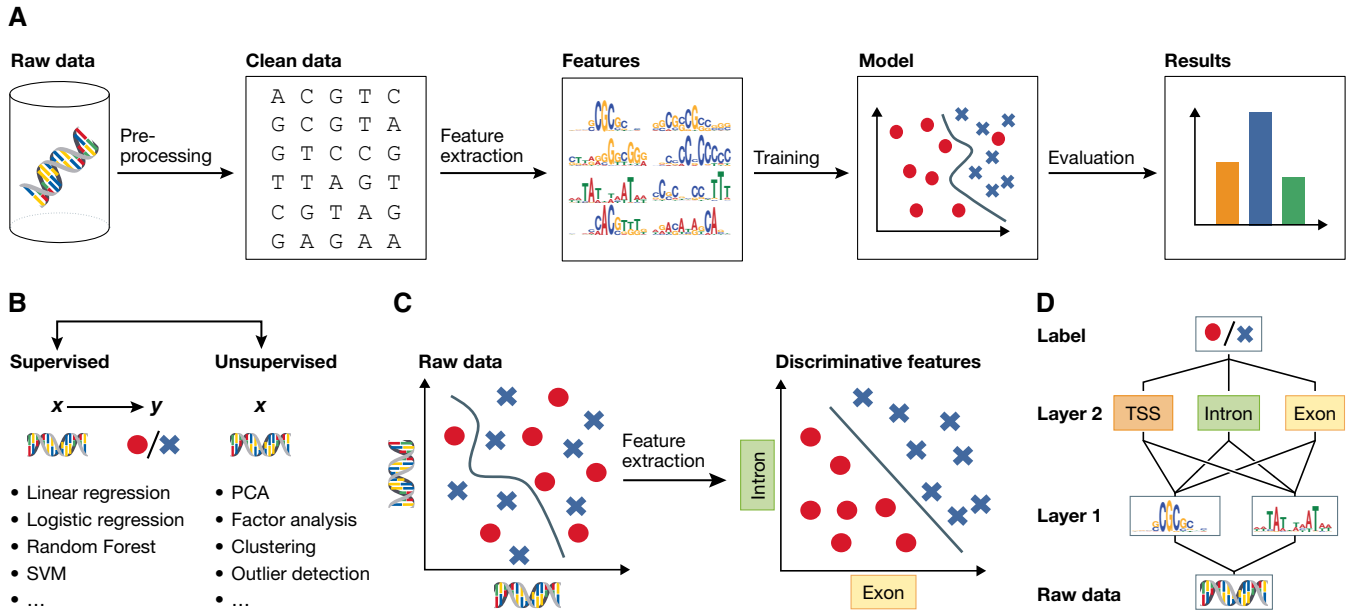


Figure 1. Machine learning and representation learning.
(A) The classical machine learning workflow can be broken down into four steps: data pre-processing, feature extraction, model learning and model evaluation. (B) Supervised machine learning methods relate input features *x* to an output label *y*, whereas unsupervised method learns factors about *x* without observed labels. (C) Raw input data are often high-dimensional and related to the corresponding label in a complicated way, which is challenging for many classical machine learning algorithms (left plot). Alternatively, higher-level features extracted using a deep model may be able to better discriminate between classes (right plot). (D) Deep networks use a hierarchical structure to learn increasingly abstract feature representations from the raw data.

process (Box 1). Deep learning is now one of the most active fields in machine learning and has been shown to improve performance in image and speech recognition (Hinton *et al*, 2012; Krizhevsky *et al*, 2012; Graves *et al*, 2013; Zeiler & Fergus, 2014; Deng & Togneri, 2015), natural language understanding (Bahdanau *et al*, 2014; Sutskever *et al*, 2014; Lipton, 2015; Xiong *et al*, 2016), and most recently, in computational biology (Eickholt & Cheng, 2013; Dahl *et al*, 2014; Leung *et al*, 2014; Sønderby & Winther, 2014; Alipanahi *et al*, 2015; Wang *et al*, 2015; Zhou & Troyanskaya, 2015; Kelley *et al*, 2016).

The potential of deep learning in high-throughput biology is clear: in principle, it allows to better exploit the availability of increasingly large and high-dimensional data sets (e.g. from DNA sequencing, RNA measurements, flow cytometry or automated microscopy) by training complex networks with multiple layers that capture their internal structure (Fig 1C and D). The learned networks discover high-level features, improve performance over traditional models, increase interpretability and provide additional understanding about the structure of the biological data.

In this review, we discuss recent and forthcoming applications of deep learning, with a focus on applications in regulatory genomics and biological image analysis. The goal of this review was not to provide comprehensive background on all technical details, which can be found in the more specialized literature (Bengio, 2012; Bengio *et al*, 2013; Deng, 2014; Schmidhuber, 2015; Goodfellow *et al*, 2016). Instead, we aimed to provide practical pointers and the necessary background to get started with deep architectures, review current software solutions and give recommendations for applying them to data. The applications we cover are deliberately broad to illustrate differences and commonalities between approaches; reviews focusing on specific domains can be found elsewhere (Park & Kellis, 2015; Gawehn *et al*, 2016; Leung *et al*, 2016; Mamoshina *et al*, 2016). Finally, we discuss both the potential and possible pitfalls of deep learning and contrast these methods to traditional machine learning and classical statistical analysis approaches.

Deep learning for regulatory genomics

Conventional approaches for regulatory genomics relate sequence variation to changes in molecular traits. One approach is to leverage variation between genetically diverse individuals to map quantitative trait loci (QTL). This principle has been applied to identify regulatory variants that affect gene expression levels (Montgomery *et al*, 2010; Pickrell *et al*, 2010), DNA methylation (Gibbs *et al*, 2010; Bell *et al*, 2011), histone marks (Grubert *et al*, 2015; Waszak *et al*, 2015) and proteome variation (Vincent *et al*, 2010; Albert *et al*, 2014; Parts *et al*, 2014; Battle *et al*, 2015) (Fig 2A). Better statistical methods have helped to increase the power to detect regulatory QTL (Kang *et al*, 2008; Stegle *et al*, 2010; Parts *et al*, 2011; Rakitsch & Stegle, 2016); however, any mapping approach is intrinsically limited to variation that is present in the training population. Thus, studying the effects of rare mutations in particular requires data sets with very large sample size.

An alternative is to train models that use variation between regions within a genome (Fig 2A). Splitting the sequence into windows centred on the trait of interest gives rise to tens of thousands of training examples for most molecular traits even when using a single individual. Even with large data sets, predicting molecular traits from DNA sequence is challenging due to multiple layers

 © 2016 The Authors

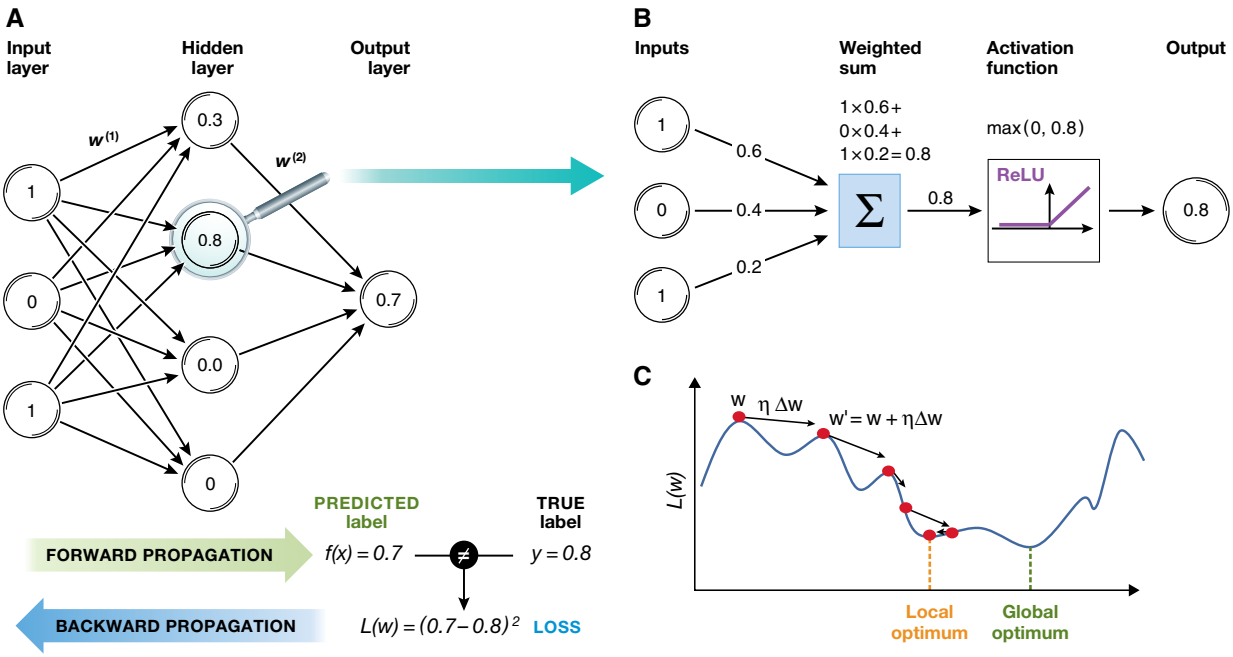
Box 1: Artificial Neural Network

An artificial neural network, initially inspired by neural networks in the brain (McCulloch & Pitts, 1943; Farley & Clark, 1954; Rosenblatt, 1958), consists of layers of interconnected compute units (neurons). The depth of a neural network corresponds to the number of hidden layers, and the width to the maximum number of neurons in one of its layers. As it became possible to train networks with larger numbers of hidden layers, artificial neural networks were rebranded to "deep networks".

In the canonical configuration, the network receives data in an input layer, which are then transformed in a nonlinear way through multiple hidden layers, before final outputs are computed in the output layer (panel A). Neurons in a hidden or output layer are connected to all neurons of the previous layer. Each neuron computes a weighted sum of its inputs and applies a nonlinear activation function to calculate its output $f(x)$ (panel B). The most popular activation function is the rectified linear unit (ReLU; panel B) that thresholds negative signals to 0 and passes through positive signal. This type of activation function allows faster learning compared to alternatives (e.g. sigmoid or tanh unit) (Glorot *et al*, 2011).

The weights $w^{(i)}$ between neurons are free parameters that capture the model's representation of the data and are learned from input/output samples. Learning minimizes a loss function $L(w)$ that measures the fit of the model output to the true label of a sample (panel A, bottom). This minimization is challenging, since the loss function is high-dimensional and non-convex, similar to a landscape with many hills and valleys (panel C). It took several decades before the *backward propagation algorithm* was first applied to compute a loss function gradient via chain rule for derivatives (Rumelhart *et al*, 1988), ultimately enabling efficient training of neural networks using stochastic gradient descent. During learning, the predicted label is compared with the true label to compute a loss for the current set of model weights. The loss is then backward propagated through the network to compute the gradients of the loss function and update (panel A). The loss function $L(w)$ is typically optimized using gradient-based descent. In each step, the current weight vector (red dot) is moved along the direction of steepest descent dw (direction arrow) by learning rate η (length of vector). Decaying the learning rate over time allows to explore different domains of the loss function by jumping over valleys at the beginning of the training (left side) and fine-tune parameters with smaller learning rates in later stages of the model training. While learning in deep neural networks remains an active area of research, existing software packages (Table 1) can already be applied without knowledge of the mathematical details involved.

Alternative architectures to such fully connected feedforward networks have been developed for specific applications, which differ in the way neurons are arranged. These include convolutional neural networks, which are widely used for modelling images (Box 2), recurrent neural networks for sequential data (Sutskever, 2013; Lipton, 2015), or restricted Boltzmann machines (Salakhutdinov & Larochelle, 2010; Hinton, 2012) and autoencoders (Hinton & Salakhutdinov, 2006; Alain *et al*, 2012; Kingma & Welling, 2013) for unsupervised learning. The choice of network architecture and other parameters can be made in a data-driven and objective way by assessing the model performance on a validation data set.



of abstraction between the effect of individual DNA variants and the trait of interest, as well as the dependence of the molecular traits on a broad sequence context and interactions with distal regulatory elements.

The value of deep neural networks in this context is twofold. First, classical machine learning methods cannot operate on the sequence directly, and thus require pre-defining features that can be extracted from the sequence based on prior knowledge (e.g. the presence or absence of single-nucleotide variants (SNVs), k-mer frequencies, motif occurrences, conservation, known regulatory variants or structural elements). Deep neural networks can help circumventing the manual extraction of features by learning them from data. Second, because of their representational richness, they can capture nonlinear dependencies in the sequence and interaction effects and span wider sequence context at multiple genomic scales. Attesting to their utility, deep neural networks have been successfully applied to predict splicing activity (Leung *et al*, 2014; Xiong *et al*, 2015), specificities of DNA- and RNA-binding proteins

© 2016 The Authors

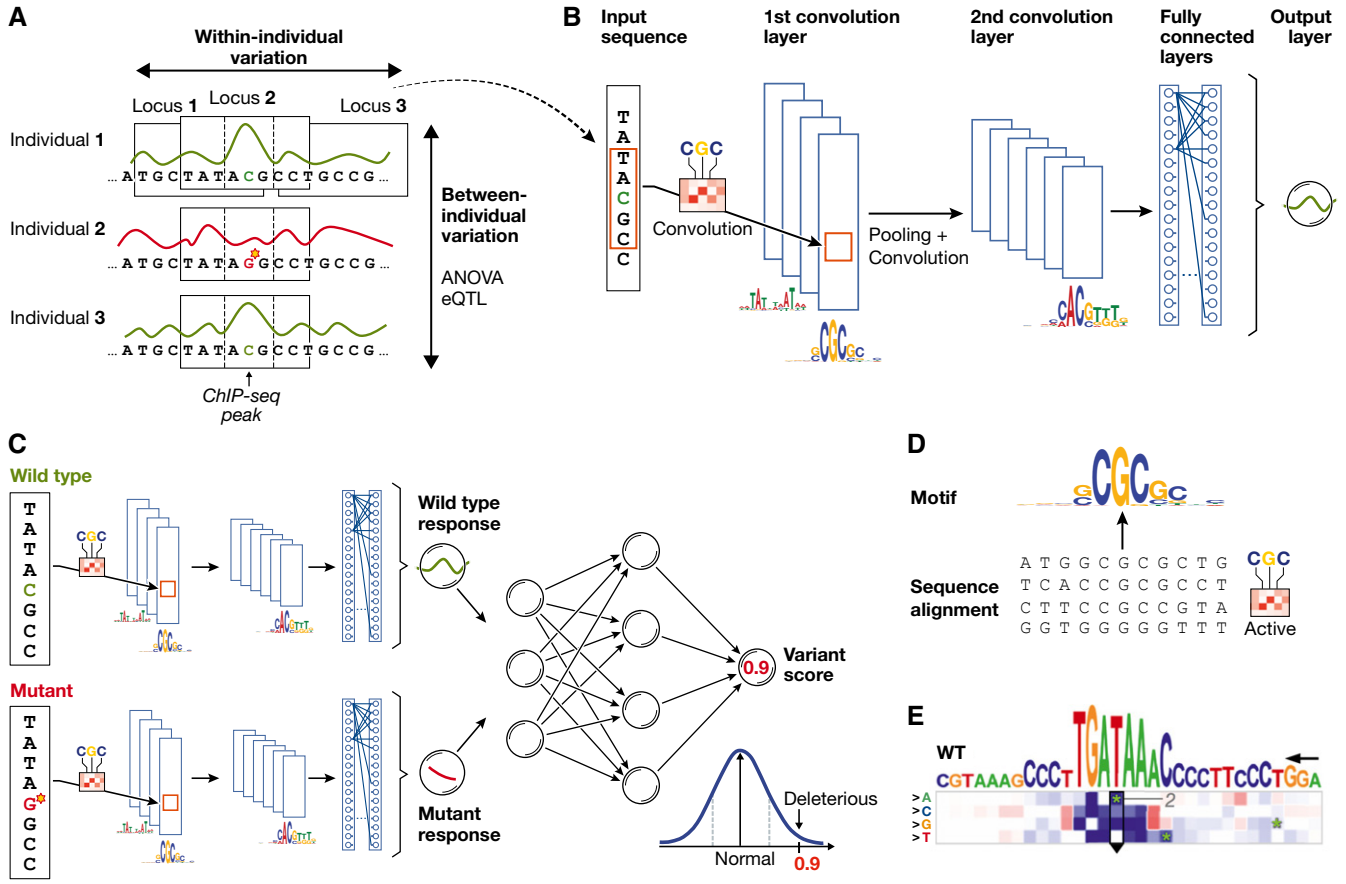


Figure 2. Principles of using neural networks for predicting molecular traits from DNA sequence.
(A) DNA sequence and the molecular response variable along the genome for three individuals. Conventional approaches in regulatory genomics consider variations between individuals, whereas deep learning allows exploiting intra-individual variations by tiling the genome into sequence DNA windows centred on individual traits, resulting in large training data sets from a single sample. (B) One-dimensional convolutional neural network for predicting a molecular trait from the raw DNA sequence in a window. Filters of the first convolutional layer (example shown on the edge) scan for motifs in the input sequence. Subsequent pooling reduces the input dimension, and additional convolutional layers can model interactions between motifs in the previous layer. (C) Response variable predicted by the neural network shown in (B) for a wild-type and mutant sequence is used as input to an additional neural network that predicts a variant score and allows to discriminate normal from deleterious variants. (D) Visualization of a convolutional filter by aligning genetic sequences that maximally activate the filter and creating a sequence motif. (E) Mutation map of a sequence window. Rows correspond to the four possible base pair substitutions, columns to sequence positions. The predicted impact of any sequence change is colour-coded. Letters on top denote the wild-type sequence with the height of each nucleotide denoting the maximum effect across mutations (figure panel adapted from Alipanahi *et al*, 2015).

(Alipanahi *et al*, 2015) or epigenetic marks and to study the effect of DNA sequence alterations (Zhou & Troyanskaya, 2015; Kelley *et al*, 2016).

Early applications of neural networks in regulatory genomics

The first successful applications of neural networks in regulatory genomics replaced a classical machine learning approach with a deep model, without changing the input features. For example, Xiong *et al* (2015) considered a fully connected feedforward neural network to predict the splicing activity of individual exons. The model was trained using more than 1,000 pre-defined features extracted from the candidate exon and adjacent introns. Despite the relatively low number of 10,700 training samples in combination with the model complexity, this method achieved substantially higher prediction accuracy of splicing activity compared to simpler approaches, and in particular was able to identify rare mutations implicated in splicing misregulation.

Convolutional designs

More recent work using convolutional neural networks (CNNs) allowed direct training on the DNA sequence, without the need to define features (Alipanahi *et al*, 2015; Zhou & Troyanskaya, 2015; Angermueller *et al*, 2016; Kelley *et al*, 2016). The CNN architecture allows to greatly reduce the number of model parameters compared to a fully connected network by applying convolutional operations to only small regions of the input space and by sharing parameters between regions. The key advantage resulting from this approach is the ability to directly train the model on larger sequence windows (Box 2; Fig 2B).

Alipanahi *et al* (2015) considered convolutional network architectures to predict specificities of DNA- and RNA-binding proteins.

 © 2016 The Authors

Box 2: Convolutional Neural Network

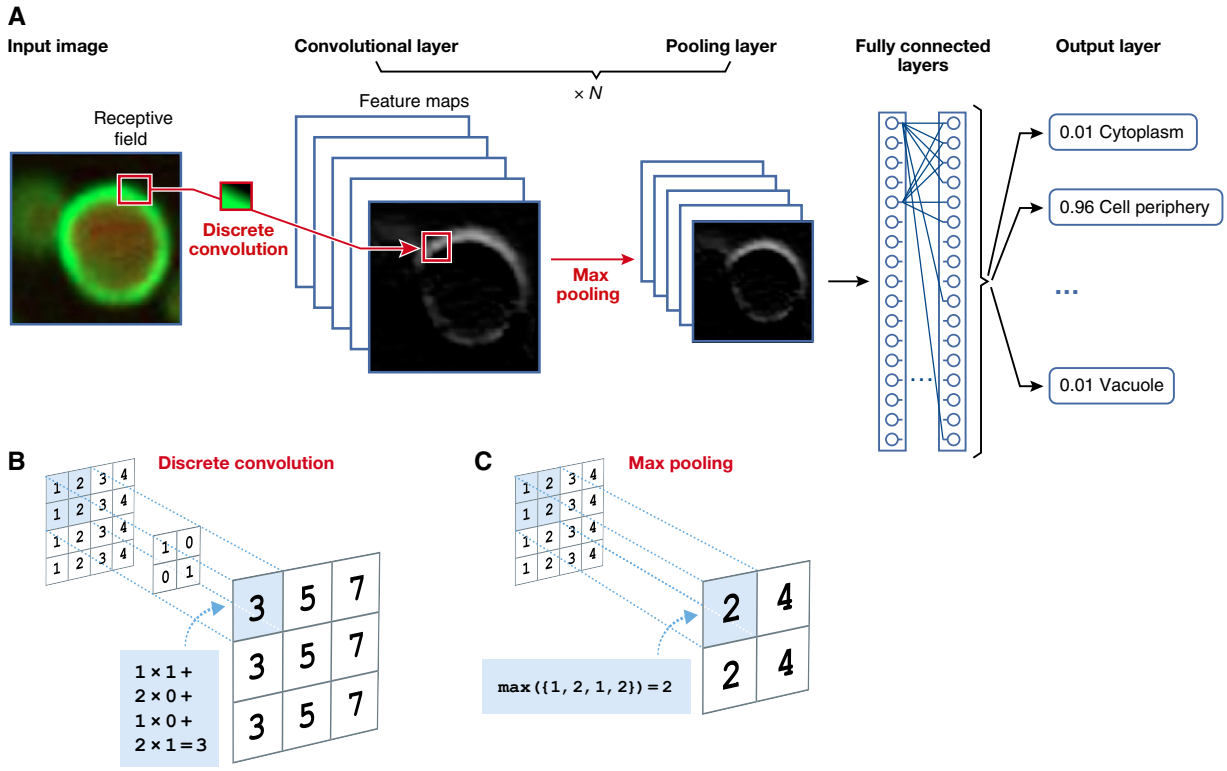
Convolutional neural networks (CNNs) were originally inspired by cognitive neuroscience and Hubel and Wiesel's seminal work on the cat's visual cortex, which was found to have simple neurons that respond to small motifs in the visual field, and complex neurons that respond to larger ones (Hubel & Wiesel, 1963, 1970).

CNNs are designed to model input data in the form of multidimensional arrays, such as two-dimensional images with three colour channels (LeCun *et al*, 1989; Jarrett *et al*, 2009; Krizhevsky *et al*, 2012; Zeiler & Fergus, 2014; He *et al*, 2015; Szegedy *et al*, 2015a) or one-dimensional genomic sequences with one channel per nucleotide (Alipanahi *et al*, 2015; Wang *et al*, 2015; Zhou & Troyanskaya, 2015; Angermueller *et al*, 2016; Kelley *et al*, 2016). The high dimensionality of these data (up to millions of pixels for high-resolution images) renders training a fully connected neural network challenging, as the number of parameters of such a model would typically exceed the number of training data to fit them. To circumvent this, CNNs make additional assumptions on the structure of the network, thereby reducing the effective number of parameters to learn.

A convolutional layer consists of multiple maps of neurons, so-called feature maps or filters, with their size being equal to the dimension of the input image (panel A). Two concepts allow reducing the number of model parameters: local connectivity and parameter sharing. First, unlike in a fully connected network, each neuron within a feature map is only connected to a local patch of neurons in the previous layer, the so-called receptive field. Second, all neurons within a given feature map share the same parameters. Hence, all neurons within a feature map scan for the same feature in the previous layer, however at different locations. Different feature maps might, for example, detect edges of different orientation in an image, or sequence motifs in a genomic sequence. The activity of a neuron is obtained by computing a discrete convolution of its receptive field, that is computing the weighted sum of input neurons, and applying an activation function (panel B).

In most applications, the exact position and frequency of features is irrelevant for the final prediction, such as recognizing objects in an image. Using this assumption, the pooling layer summarizes adjacent neurons by computing, for example, the maximum or average over their activity, resulting in a smoother representation of feature activities (panel C). By applying the same pooling operation to small image patches that are shifted by more than one pixel, the input image is effectively down-sampled, thereby further reducing the number of model parameters.

A CNN typically consists of multiple convolutional and pooling layers, which allows learning more and more abstract features at increasing scales from small edges, to object parts, and finally entire objects. One or more fully connected layers can follow the last pooling layer (panel A). Model hyper-parameters such as the number of convolutional layers, number of feature maps or the size of receptive fields are application-dependent and should be strictly selected on a validation data set.



Their *DeepBind* model outperformed existing methods, was able to recover known and novel sequence motifs, and could quantify the effect of sequence alterations and identify functional SNVs. A key innovation that enabled training the model directly on the raw DNA sequence was the application of a one-dimensional convolutional layer. Intuitively, the neurons in the convolutional layer scan for motif sequences and combinations thereof, similar to conventional position weight matrices (Stormo *et al*, 1982). The learning signal from deeper layers informs the convolutional layer which motifs are the most relevant. The motifs recovered by the model can then be visualized as heatmaps or sequence logos (Fig 2D).

© 2016 The Authors

In silico prediction of mutation effects

An important application of deep neural networks trained on the raw DNA sequence is to predict the effect of mutations *in silico*. Such model-based assessments of the effect of sequence changes complement methods based on QTL mapping, and can in particular help to uncover regulatory effects of rare SNVs or to fine-map likely causal genes. An intuitive approach for visualizing such predicted regulatory effects is mutation maps (Alipanahi *et al*, 2015), whereby the effect of all possible mutations for a given input sequence is represented in a matrix view (Fig 2E). The authors could further reliably identify deleterious SNVs by training an additional neural network with predicted binding scores for a wild-type and mutant sequence (Fig 2C).

Joint prediction of multiple traits and further extensions

Following their initial successes, convolutional architectures have been extended and applied to a range of tasks in regulatory genomics. For example, Zhou and Troyanskaya (2015) considered these architectures to predict chromatin marks from DNA sequence. The authors observed that the size of the input sequence window is a major determinant of model performance, where larger windows (now up to 1 kb) coupled with multiple convolutional layers enabled capturing sequence features at different genomic length scales. A second innovation was to use neural network architectures with multiple output variables (so-called multitask neural networks) to predict multiple chromatin states in parallel. Multitask architectures allow learning shared features between outputs, thereby improving generalization performance, and markedly reducing the computational cost of model training compared to learning independent models for each trait (Dahl *et al*, 2014).

In a similar vein, Kelley *et al* (2016) developed the open-source deep learning framework *Basset*, to predict DNase I hypersensitivity across multiple cell types and to quantify the effect of SNVs on chromatin accessibility. Again, the model improved prediction performance compared to conventional methods and was able to retrieve both known and novel sequence motifs that are associated with DNase I hypersensitivity. A related architecture has also been considered by Angermueller *et al* to predict DNA methylation states in single-cell bisulphite sequencing studies (Angermueller *et al*, 2016). This approach combined convolutional architectures to detect informative DNA sequence motifs with additional features derived from neighbouring CpG sites, thereby accounting for methylation context. Most recently, Koh, Pierson and Kundaje applied CNNs to de-noise genomewide chromatin immunoprecipitation followed by sequencing data in order to obtain a more accurate prevalence estimate for different chromatin marks (Koh *et al*, 2016).

At present, CNNs are among the most widely used architectures to extract features from fixed-size DNA sequence windows. However, alternative architectures could also be considered. For example, recurrent neural networks (RNNs) are suited to model sequential data (Lipton, 2015) and have been applied for modelling natural language and speech (Hinton *et al*, 2012; Graves *et al*, 2013; Sutskever *et al*, 2014; Che *et al*, 2015; Deng & Togneri, 2015; Xiong *et al*, 2016), protein sequences (Agathocleous *et al*, 2010; Sønderby & Winther, 2014), clinical medical data (Che *et al*, 2015; Lipton *et al*, 2015) and to a limited extent DNA sequences (Xu *et al*, 2007; Lee *et al*, 2015). RNNs are appealing for applications in regulatory genomics, because they allow modelling sequences of variable length, and to capture long-range interactions within the sequence and across multiple outputs. However, at present, RNNs are more difficult to train than CNNs, and additional work is needed to better understand the settings where one should be preferred over the other.

Complementary to supervised methods, unsupervised deep learning architectures learn low-dimensional feature representations from high-dimensional unlabelled data, similarly to classical principal component analysis or factor analysis, but using a nonlinear model. Examples of such approaches are stacked autoencoders (Vincent *et al*, 2010), restricted Boltzmann machines and deep belief networks (Hinton *et al*, 2006). The learnt features can be used to visualize data or as input for classical supervised learning tasks. For example, sparse autoencoders have been applied to classify cancer cases using gene expression profiles (Fakoor *et al*, 2013) or to predict protein backbones (Lyons *et al*, 2014). Restricted Boltzmann machines can also be used for unsupervised pre-training of deep networks to subsequently train supervised models of protein secondary structures (Spencer *et al*, 2015), disordered protein regions (Eickholt & Cheng, 2013) or amino acid contacts (Eickholt & Cheng, 2012). Skip-gram neural networks have been applied to learn low-dimensional representations of protein sequences and improve protein classification (Asgari & Mofrad, 2015). In general, unsupervised models are a powerful approach if large quantities of unlabelled data are available to pre-train complex models. Once trained, these models can help to improve performance on classification tasks, for which smaller numbers of labelled examples are typically available.

Deep learning for biological image analysis

Historically, perhaps the most important successes of deep neural networks have been in image analysis. Deep architectures trained on millions of photographs can famously detect objects in pictures better than humans do (He *et al*, 2015). All current state-of-the-art models in image classification, object detection, image retrieval and semantic segmentation make use of neural networks.

The convolutional neural network (Box 2) is the most common network architecture for image analysis. Briefly, a CNN performs pattern matching (convolution) and aggregation (pooling) operations (Box 2). At a pixel level, the convolution operation scans the image with a given pattern and calculates the strength of the match for every position. Pooling determines the presence of the pattern in a region, for example by calculating the maximum pattern match in smaller patches (max-pooling), thereby aggregating region information into a single number. The successive application of convolution and pooling operations is at the core of most network architectures used in image analysis (Box 2).

First applications in computational biology—pixel-level classification

The early applications of deep networks for biological images focused on pixel-level tasks, with additional models building on the network outputs. For example, Ning *et al* (2005) applied

 © 2016 The Authors

convolutional neural networks in a study that predicted abnormal development in *C. elegans* embryo images. They trained a CNN on 40 × 40 pixel patches to classify the centre pixel to cell wall, cytoplasm, nucleus membrane, nucleus or outside medium, using three convolutional and pooling layers, followed by a fully connected output layer. The model predictions were then fed into an energy-based model for further analysis. CNNs have outperformed standard methods, for example Markov random fields and conditional random fields (Li, 2009) in such raw data analysis tasks, for example restoring noisy neural circuitry images (Jain *et al*, 2007).

Adding layers allows moving from clearing up pixel noise to modelling more abstract image features. Ciresan *et al* (2013) used five convolutional and pooling layers, followed by two fully connected layers, to find mitosis in breast histology images. This model won the mitosis detection challenge at the International Conference of Pattern Recognition 2012, outperforming competitors by a substantial margin. The same approach was also used to segment neuronal structures in electron microscopy images, classifying each pixel as membrane or non-membrane (Ciresan *et al*, 2012). In these applications, while the CNNs were trained in an end-to-end manner, additional post-processing was required to obtain class probabilities from the outputs for new images.

Successive pooling operations lose information on localization, as only summaries are retained from larger and larger regions. To avoid this, skip links can be added to carry information from early, fine-grained layers forward to deeper ones. The currently best-performing pixel-level classification method for neuronal structures (*U-Net;* Ronneberger *et al*, 2015) employs an architecture in which neurons take inputs from lower layers to localize high-resolution features, as well as to overcome the arbitrary choice of context size.

Analysis of whole cells, cell populations and tissues

In many cases, pixel-level predictions are not required. For example, Xu *et al* directly classified colon histopathology images into cancerous and non-cancerous, finding that supervised feature learning with deep networks was superior to using handcrafted features (Xu *et al*, 2014). Pärnamaa and Parts used CNNs to classify pre-segmented image patches of individual yeast cells carrying a fluorescent protein to different subcellular localization patterns (Pärnamaa & Parts, 2016). Again, deep networks outperformed methods based on traditional features. Further, Kraus *et al* combined the segmentation and classification tasks into a single architecture that can be learned end-to-end and applied the model to full resolution yeast microscopy images (Kraus *et al*, 2015). This approach allowed classifying entire images without performing segmentation as a pre-processing step. CNNs have even been applied to count bacterial colonies in agar plates (Ferrari *et al*, 2015). Since the early denoising applications on the pixel level, the field has been moving towards end-to-end image analysis pipelines that make use of large bioimage data sets, and the representational power of CNNs.

Reusing trained models

Training convolutional neural networks requires large data sets. While biological data acquisition can be expensive, this does not mean that deep neural networks cannot be used when millions of images are not available. Regardless of image source, lower levels of the network tend to capture similar signal (edges, blobs) that are not specific to the training data and the application, but instead recur in perceptual tasks in general. Thus, convolutional neural networks can reuse pictures from a similar domain to help with learning, or even be pre-trained on other data, thereby requiring fewer images to fine-tune the model for the task of interest. Indeed, Donahue *et al* (2013) and Razavian *et al* (2014) showed that features learned from millions of images to classify objects, can successfully be used in image retrieval, detection or classification on new domains where only hundreds of images are labelled. The effectiveness of such an approach depends on the similarity between the training data and the new domain (Yosinski *et al*, 2014).

The concept of transferring model parameters has also been successful in bioimage analysis. For example, Zhang *et al* (2015) showed that features learned from natural images can be transferred to biological data, improving the prediction of *Drosophila melanogaster* developmental stages from *in situ* hybridization images. The model was first pre-trained on data from the ImageNet (Russakovsky *et al*, 2015), an open corpus of more than one million diverse images, to extract rich features at different scales. Xie *et al* (2015) further used synthetic images to train a CNN for automatic cell counting in microscopy images. We expect that network repositories that host pre-trained models will emerge for biological image analysis; such efforts already exist for general image processing tasks (see learning section below). These trained models could be downloaded and used as feature extractors (Fig 3), or further fine-tuned and adapted to a particular task on small-scale data.

Interpreting and visualizing convolutional networks

Convolutional neural networks have been successful across many domains. In interpreting their performance, it is useful to understand the features they capture.

Visualizing input weights

One way to understand what a particular neuron represents is to look for inputs that maximally activate it. Under some mathematical constraints, these patterns are proportional to the incoming weights (see also Box 1). Krizhevsky *et al* visualized weights in the first convolutional layer (Krizhevsky *et al*, 2012) and found that these maximally activating patterns correspond to colour blobs, edges at different orientations and Gabor-like filters (Fig 4). Gabor filters are widely used pre-defined features in image analysis; neural networks rediscover them in a data-driven way as a useful component of the image model. Higher layer weights can be visualized as well, but as the inputs are not pixels, their weights are more difficult to interpret.

Finding images that maximize neuron activity

To understand the deeper layers in terms of input pixels, Girshick *et al* (2014) retrieved and Simonyan *et al* (2013) generated images that maximize the output of individual neurons (Fig 4). While this approach yields no explicit representation, it can provide an overview of the type of features that differentiate images with large neuron activity from all others. Such visualizations tend to show that second-layer features combine edges from the first layer,

© 2016 The Authors

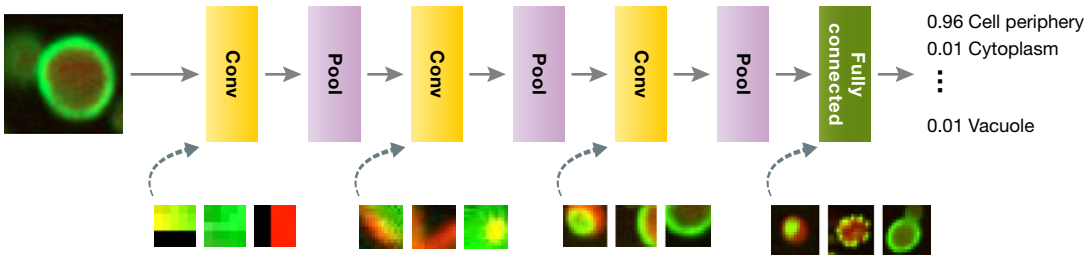


Figure 3. Convolution and pooling operators are stacked, thereby creating a deep network for image analysis.
In standard applications, convolution layers are followed by a pooling layer (Box 2). In this example, the lowest level convolutional units operate on 3 × 3 patches, but deeper ones use and capture information from larger regions. These convolutional pattern-matching layers are followed by one or multiple fully connected layers to learn which features are most informative for classification. For each layer with learnable weights, three example images that maximize some neuron output are shown.

thereby detecting corners and angles; deeper layer neurons activate for specific object parts (e.g. noses, eyes); and the deepest layers detect whole objects (e.g. faces, cars). It is complicated to hand-engineer features that look specifically for noses, eyes or faces, but neural networks can learn these features solely from input–output examples.

Hiding important image parts

To understand which image parts are important for determining the value of each feature, Zeiler and Fergus (2014) occluded images with smaller grey boxes. The parts that are most influential will drastically change the feature value when occluded. In a similar vein, Simonyan *et al* (2013) and Springenberg *et al* (2014) visualized which individual pixels make the most difference in the feature, and Bach, Binder and colleagues developed pixel relevance for individual classification decisions in a more general framework (Bach *et al*, 2015). This information can also be used for object localization or segmentation, as the sensitive image pixels usually correctly correspond to the true object. Kraus *et al* (2015) used this idea to effectively localize cells in large microscopy images.

Visualizing similar inputs in two dimensions

Visualizing the CNN representations can help gauge what inputs get mapped to similar feature vectors, and hence understand what the model has learned. Donahue *et al* (2013) projected CNN features into two dimensions to show that each subsequent layer transforms data to be more and more separable by a linear classifier. In general, different CNN visualization methods show that higher layer features are more specific to the learning task, while low-level features tend to capture general aspects of images, such as edges and corners.

Off-the-shelf tools and practical considerations

Deep learning frameworks

Deep learning frameworks have been developed to easily build neural networks from existing modules on a high level. The most popular ones are Caffe (Jia *et al*, 2014), Theano (Bastien *et al*, 2012), Torch7 (Collobert *et al*, 2011) and TensorFlow (Abadi *et al*, 2016; Rampasek & Goldenberg, 2016) (Table 1), which differ in modularity, ease of use and the way models are defined and trained.

Caffe (Jia *et al*, 2014) is developed by the Berkeley Vision and Learning Center and is written in C++. The network architecture is specified in a configuration file and models can be trained and used via command line, without writing code at all. Additionally, Python and MATLAB interfaces are available. Caffe offers one of the most efficient implementations for CNNs and provides multiple pre-trained models for image recognition, making it well suited for computer vision tasks. As a downside, custom models need to be implemented in C++, which can be difficult. Additionally, Caffe is not optimized for recurrent architectures.

Theano (Bastien *et al*, 2012; Team *et al*, 2016) is developed and maintained by the University of Montreal and written in Python and C++. Model definitions follow a declarative instead of an imperative programing paradigm, which means that the user specifies what needs to be done, not in which order. A neural network is declared as a computational graph, which is then compiled to native code and executed. This design allows Theano to optimize computational steps and to automatically derive gradients—one of its main strengths. Consequently, Theano is well suited for building custom models and offers particularly efficient implementations for RNNs. Software wrappers such as Keras (https://github.com/fchollet/keras) or Lasagne (https://github.com/Lasagne/Lasagne) provide additional abstraction and allow building networks from existing components, and reusing pre-trained networks. The major drawback of Theano is frequently long compile times when building larger models.

Torch7 (Collobert *et al*, 2011) was initially developed at the University of New York and is based on the scripting language LuaJIT. Networks can be easily built by stacking existing modules and are not compiled, hence making it more suited for fast prototyping than Theano. Torch7 offers an efficient CNN implementation and access to a range of pre-trained models. A possible downside is the need of the user to be familiar with the LuaJIT scripting language. Also, LuaJIT is less suited for building custom recurrent networks.

TensorFlow (Abadi *et al*, 2016) is the most recent deep learning framework developed by Google. The software is written in C++ and offers interfaces to Python. Similar to Theano, a neural network is declared as a computational graph, which is optimized during compilation. However, the shorter compile time makes it more suited for prototyping. A key strength of TensorFlow is native support for parallelization across different devices, including CPUs and GPUs, and using multiple compute nodes on a cluster. The accompanying tool TensorBoard allows to conveniently visualize networks in a web browser and to monitor training progress, for example learning curves or parameter updates. At present,

 © 2016 The Authors

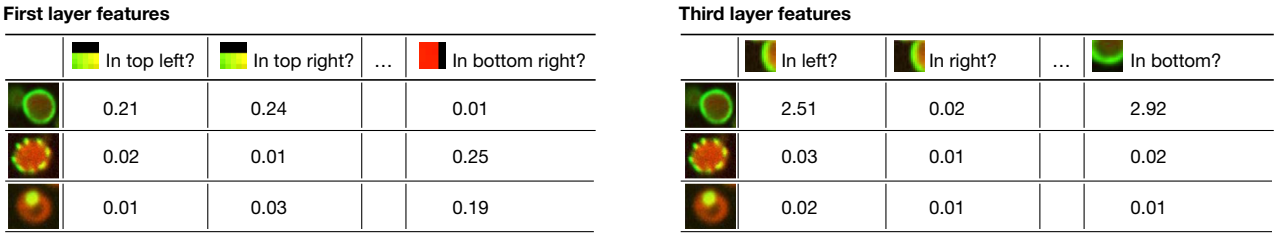


Figure 4. A pre-trained network can be used as a generic feature extractor.
Feeding input into the first layer (left) gives a low-level feature representation in terms of patterns (left to right) present in smaller patches in every cell (top to bottom). Neuron activations extracted from deeper layers (right) give rise to more abstract features that capture information from a larger segment of the image.

TensorFlow provides the most efficient implementation for RNNs. The software is recent and under active development; hence, only few pre-trained models are currently available.

Data preparation

Training data are key for every machine learning application. Since more data with informative features usually result in better performance, effort should be spent on collecting, labelling, cleaning and normalizing data.

Required data set sizes

Most of the successful applications of deep learning have been in supervised learning settings, where sufficient labelled training samples are available to fit complex models. As a rule of thumb, the number of training samples should be at least as high as the number of model parameters, although special architectures and model regularization can help to avoid overfitting if training data are scarce (Bengio, 2012).

Central problems in regulatory genomics, for example predicting molecular traits from genotype, are limited in the number of training instances; hundreds to at most tens of thousands of training examples are typical. The strategy of considering sequence windows centred on the trait of interest (e.g. splice site, transcription factor binding site or epigenetic marks; see Fig 2A) is now a widely used approach and helps increasing the number of input–output pairs from a single individual.

In image analysis, data can be abundant, but manually curated and labelled training examples are typically difficult to obtain. In such instances, the training set can be augmented by scaling, rotating or cropping the existing images, an approach that also enhances robustness (Krizhevsky *et al*, 2012). Another strategy is to reuse a network that was pre-trained on a large data set for image recognition [e.g. AlexNet (Krizhevsky *et al*, 2012), VGG (Simonyan & Zisserman, 2014), GoogleNet (Szegedy *et al*, 2015b) or ResNet (He *et al*, 2015)], and to fine-tune its parameters on the data set of interest (e.g. microscopy images for a particular segmentation task). Such an approach exploits the fact that different data sets share important characteristics and features, such as edges or curves, which can be transferred between them. Caffe, Lasagne, Torch and to a limited extend TensorFlow provide repositories with pre-trained models.

Partitioning data into training, validation and test sets

Machine learning models need to be trained, selected and tested on independent data sets to avoid overfitting and assure that the model will generalize to unseen data. Holdout validation, partitioning the data into a training, validation and test sets, is the standard for deep neural networks (Fig 5C). The training set is used to learn models with different hyper-parameters, which are then assessed on the validation set. The model with best performance, for example prediction accuracy or mean-squared error, is selected and further evaluated on the test set to quantify the performance on unseen data and for comparison to other methods. Typical data set proportions are 60% for training, 10% for validation and 30% for model testing. If the data set is small, k-fold cross-validation or bootstrapping can be used instead (Hastie *et al*, 2005).

Normalization of raw data

Appropriate choices for data normalization can help to accelerate training and the identification of a good local minimum.

Categorical features such as DNA nucleotides first need to be encoded numerically. They are typically represented as binary vectors with all but one entry set to zero, which indicates the category (*one-hot coding*). For example, DNA nucleotides (categories)

Table 1. Overview of existing deep learning frameworks, comparing four widely used software solutions.

	Caffe	Theano	Torch7	TensorFlow
Core language	C++	Python, C++	LuaJIT	C++
Interfaces	Python, Matlab	Python	C	Python
Wrappers		Lasagne, Keras, sklearn-theano		Keras, Pretty Tensor, Scikit Flow
Programming paradigm	Imperative	Declarative	Imperative	Declarative
Well suited for	CNNs, Reusing existing models, Computer vision	Custom models, RNNs	Custom models, CNNs, Reusing existing models	Custom models, Parallelization, RNNs

© 2016 The Authors

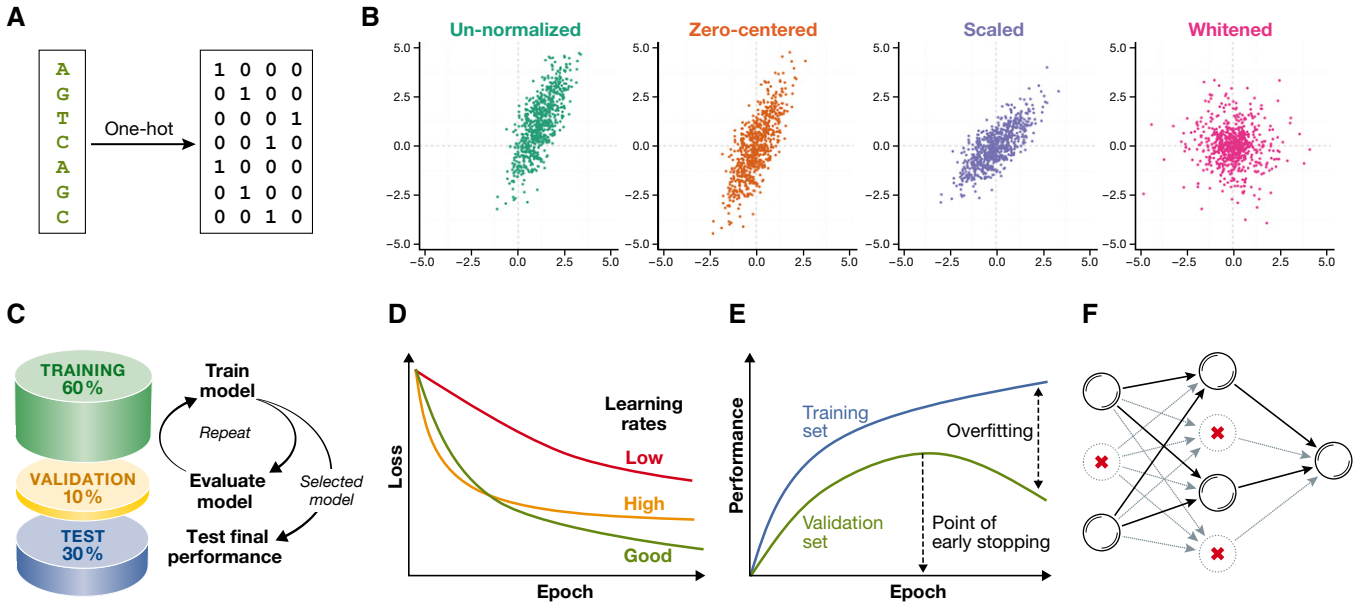


Figure 5. Data normalization for and pre-processing for deep neural networks.
(A) DNA sequence one-hot encoded as binary vectors using codes A = 1 0 0 0, G = 0 1 0 0, C = 0 0 1 0 and T = 0 0 0 1. (B) Continuous data (green) after zero-centring (orange), scaling to unit variance (blue) and whiting (purple). (C) Holdout validation partitions the full data set randomly into training (~60%), validation (~10%) and test set (~30%). Models are trained with different hyper-parameters on the training set, from which the model with the highest performance on the validation set is selected. The generalization performance of the model is assessed and compared with other machine learning methods on the test set. (D) The shape of the learning curve indicates if the learning rate is too low (red, shallow decay), too high (orange, steep decay followed by saturation) or appropriate for a particular learning task (green, gradual decay). (E) Large differences in the model performance on the training set (blue) and validation set (green) indicate overfitting. Stopping the training as soon as the validation set performance starts to drop (early stopping) can prevent overfitting. (F) Illustration of the dropout regularization. Shown is a feedforward neural network after randomly dropping out neurons (crossed out), which reduces the sensitivity of neurons to neurons in the previous layer due to non-existent inputs (greyed edges).

are commonly encoded as A = (1 0 0 0), G = (0 1 0 0), C = (0 0 1 0) and T = (0 0 0 1) (Fig 5A). A DNA sequence can then be represented as a binary string by concatenating the encoding nucleotides, and treating each nucleotide as an independent input feature of a feedforward neural network. In a CNN, the four bits of each encoded base are commonly considered analogously to colour channels of an image to preserve the entity of a nucleotide.

Numerical features are typically zero-centred by subtracting their mean value. Image pixels are usually not zero-centred individually, but jointly by subtracting the mean pixel intensity per colour channel. An additional common normalization step is to standardize features to unit variance. Whiting can be used to decorrelate features (Fig 5B), but can be computationally involved, since it requires computing the feature covariance matrix (Hastie *et al*, 2005). If the distribution of features is skewed due to a few extreme values, log transformations or similar processing steps may be appropriate. Validation and test data need to be normalized consistently with the training data. For example, features of the validation data need to be zero-centred by subtracting the mean computed on the training data, not on the validation data.

Model building

Choice of model architecture

After preparing the data, design choices about the model architectures need to be made. The default architecture is a feedforward neural network with fully connected hidden layers, which is an appropriate starting point for many problems. Convolutional architectures are well suited for multi- and high-dimensional data, such as two-dimensional images or abundant genomic data. Recurrent neural networks can capture long-range dependencies in sequential data of varying lengths, such as text, protein or DNA sequences. More sophisticated models can be built by combining different architectures. To describe the content of an image, for example, a CNN can be combined with an RNN, where the CNN encodes the image and the RNN generates the corresponding image description (Vinyals *et al*, 2015; Xu *et al*, 2015). Most deep learning frameworks provide modules for different architectures and their combinations.

Determining the number of neurons in a network

The optimal number of hidden layers and hidden units is problem-dependent and should be optimized on a validation set. One common heuristic is to maximize the number of layers and units without overfitting the data. More layers and units increase the number of representable functions and local optima, and empirical evidence shows that it makes finding a good local optimum less sensitive to weight initialization (Dauphin *et al*, 2014).

Model training

The goal of model training is to find parameters w that minimize an objective function $L(w)$, which measures the fit between the predictions the model parameterized by w and the actual observations.

 © 2016 The Authors

The most common objective functions are the cross-entropy for classification and mean-squared error for regression. Minimizing $L(w)$ is challenging since it is high-dimensional and non-convex (Fig 5C); see also Box 1 and Fig 2.

Stochastic gradient descent

Stochastic gradient descent is widely used to train deep models. Starting from an initial set of parameters w_0, the gradient dw of L with respect to w is computed for a random batch of only few, for example 128, training samples. dw points to the direction of steepest descent, towards which w is updated with step size eta, the learning rate (Fig 1C). At each step, the parameters are updated into the direction of steepest descent until a minimum is reached, analogously to a ball running down a hill to a valley (Bengio, 2012). The training performance strongly depends on parameter initialization, learning rate and batch size.

Parameter initialization

In general, model parameters should be initialized randomly to avoid local optima determined by a fixed initialization. Starting points for model parameters can be sampled independently from a normal distribution with small variance, or more commonly from a normal distribution with its variance scaled inversely by the number of hidden units in the input layer (Glorot & Bengio, 2010; He *et al*, 2015).

Learning rate and batch size

The learning rate and batch size of stochastic gradient descent need to be chosen with care, since they can strongly impact training speed and model performance. Different learning rates are usually explored on a logarithmic scale such as 0.1, 0.01 or 0.001, with 0.01 as the recommended default value (Bengio, 2012). A batch size of 128 training samples is suitable for most applications. The batch size can be increased to speed up training or decreased to reduce memory usage, which can be important for training complex models on memory-limited GPUs. The optimum learning rate and batch size are connected, with larger batch sizes typically requiring smaller learning rates.

Learning rate decay

The learning rate can be gradually reduced during training, which is based on the idea that larger steps may be helpful in early training stages in order to overcome possible local optima, whereas smaller step sizes allow exploring narrow parameter regions of the loss function in advanced stages of training. Common approaches include to linearly reduce the learning rate by a constant factor such as 0.5 after the validation loss stops improving, or exponentially after every training iteration or epoch (Bengio, 2012; Gawehn *et al*, 2016).

Momentum

Vanilla stochastic gradient descent can be extended by "momentum", which usually improves training (Sutskever *et al*, 2013). Instead of updating the current parameter vector w_t at time t by the gradient vector dw_{t+1} directly, a fraction of the previous update is added to the current one. With momentum rate v, weights are updated by a momentum vector $m_{t+1} = v \cdot m_t - \in dW_{t+1}$. This approach can help to take larger steps in directions where gradients point consistently, and therefore speed up the convergence. The momentum rate v can be set between [0, 1], and a typical value is 0.9. Nesterov momentum (Nesterov, 1983, 2013) is a special form of the same concept, which sometimes provides additional advantages.

Per-parameter adaptive learning rate methods

To reduce the sensitivity to the specific choice of the learning rate, adaptive learning rate methods, such as RMSprop, Adagrad (Srivastava *et al*, 2014) and Adam (Kingma & Ba, 2014), have been developed in order to appropriately adapt the learning rate per parameter during training. The most recent method, Adam, combines the strengths of previous methods RMSprop and Adagrad and is generally recommended for many applications.

Batch normalization

Batch normalization (Ioffe & Szegedy, 2015) is a recently described approach to reduce the dependency of training to the parameter initialization, speed up training and reduce overfitting. It is easy to implement, has marginal additional compute costs and has hence become common practice. Batch normalization zero centres and normalizes data not only at the input layer, but also at hidden layers before the activation function. This approach allows using higher learning rates and hence also accelerates training.

Analysing the learning curve

To validate the learning process, the loss should be monitored as a function of the number of training epochs, that is the number of times the full training set has been traversed (Fig 5D). If the learning curve decreases slowly, the learning rate may be too small and should be increased. If the loss decreases steeply at the beginning but saturates quickly, the learning rate may be too high. Extreme learning rates can result in an increasing or fluctuating learning curve (Bengio, 2012).

Monitoring training and validation performance

In parallel with the training loss, it is recommended to monitor the target performance such as the accuracy for both the training and validation set during training (Fig 5E). A low or decreasing validation performance relative to the training performance indicates overfitting (Bengio, 2012).

Avoiding overfitting

Deep neural networks are notoriously difficult to train, and overfitting to data is a major challenge, since they are nonlinear and have many parameters. Overfitting results from a too complex model relative to the size of the training set, and can thus be reduced by decreasing the model complexity, for example the number of hidden layers and units, or by increasing the size of the training set, for example via data augmentation. The following training guidelines can help to avoid overfitting.

Dropout (Srivastava *et al*, 2014) is the most common regularization technique and often one of the key ingredients to train deep models. Here, the activation of some neurons is randomly set to zero ("dropped out") during training in each forward pass, which intuitively results in an ensemble of different networks whose

© 2016 The Authors

predictions are averaged (Fig 5E). The dropout rate corresponds to the probability that a neuron is dropped out, where 0.5 is a sensible default value. In addition to dropping out hidden units, input units can be dropped, however usually at a lower rate. Dropout is often combined with regularizing the magnitude or parameter values by the L2 norm, and less commonly the L1 norm.

Another popular regularization method is "early stopping". Here, training is stopped as soon as the validation performance starts to saturate or deteriorate, and the parameters with the best performance on the validation set chosen.

Layerwise pre-training (Bengio *et al*, 2007; Salakhutdinov & Hinton, 2012) should be considered if the model overfits despite the mentioned regularization techniques. Instead of training the entire network at once, layers are first pre-trained unsupervised using autoencoders or restricted Boltzmann machines. Afterwards, the entire network is fine-tuned using the actual supervised learning objective.

Hyper-parameter optimization

Table 2 summarizes recommendations and starting points for the most common hyper-parameters, excluding architecture-dependent hyper-parameters such as the size and number of filters of a CNN. Since the best hyper-parameter configuration is data- and application-dependent, models with different configurations should be trained and their performance be evaluated on a validation set. As the number of configurations grows exponentially with the number of hyper-parameters, trying all of them is impossible in practice (Bengio, 2012). It is therefore recommended to optimize the most important hyper-parameters such as the learning rate, batch size or length of convolutional filters independently via line search, which is exploring different values while keeping all other hyper-parameters constant. The refined hyper-parameter space can then be further explored by random sampling, and settings with the best performance on the validation set are chosen. Frameworks such as Spearmint (Snoek *et al*, 2012), Hyperopt (Bergstra & Cox, 2013) or SMAC (Hutter *et al*, 2011) allow to automatically explore the hyper-parameter space using Bayesian optimization. However, although conceptually more powerful, they are at present more difficult to apply and parallelize than random sampling.

Table 2. Central parameters of a neural network and recommended settings.

Name	Range	Default value
Learning rate	0.1, 0.01, 0.001, 0.0001	0.01
Batch size	64, 128, 256	128
Momentum rate	0.8, 0.9, 0.95	0.9
Weight initialization	Normal, Uniform, Glorot uniform	Glorot uniform
Per-parameter adaptive learning rate methods	RMSprop, Adagrad, Adadelta, Adam	Adam
Batch normalization	Yes, no	Yes
Learning rate decay	None, linear, exponential	Linear (rate 0.5)
Activation function	Sigmoid, Tanh, ReLU, Softmax	ReLU
Dropout rate	0.1, 0.25, 0.5, 0.75	0.5
L1, L2 regularization	0, 0.01, 0.001	

Training on GPUs

Training neural networks is more time-consuming compared to shallow models and can take hours, days or even weeks, depending on the size of training set and model architecture. Training on GPUs can considerably reduce the training time (commonly by tenfold or more) and is therefore crucial for evaluating multiple models efficiently. The reason for this speedup is that learning deep networks requires large numbers of matrix multiplications, which can be parallelized efficiently on GPUs. All state-of-the-art deep learning frameworks provide support to train models on either CPUs or GPUs without requiring any knowledge about GPU programming. On desktop machines, the local GPU card can often be used if the framework supports the specific brand. Alternatively, commercial providers provide GPU cloud compute clusters.

Pitfalls

No single method is universally applicable, and the choice of whether and how to use deep learning approaches will be problem-specific. Conventional analysis approaches will remain valid and have advantages when data are scarce or if the aim is to assess statistical significance, which is currently difficult using deep learning methods. Another limitation of deep learning is the increased training complexity, which applies both to model design and the required compute environment.

Conclusion

Deep learning methods are a powerful complement to classical machine learning tools and other analysis strategies. Already, these approaches have found use in a number of applications in computational biology, including regulatory genomics and image analysis. The first publicly available software frameworks have helped to reduce the overhead of model development and provided a rich, accessible toolbox to practitioners. We expect that continued improvement of software infrastructure will make deep learning applicable to a growing range of biological problems.

Acknowledgements

OS and CA were funded by the European Molecular Biology Laboratory. TP was supported by the European Regional Development Fund through the BioMedIT project, and Estonian Research Council (IUT34-4). LP was supported by the Wellcome Trust and Estonian Research Council (IUT34-4). OS was supported by the European Research Council (agreement N635290).

Conflict of interest

The authors declare that they have no conflict of interest.

 © 2016 The Authors

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Josofowicz R, Kaiser L, Kudlur M, Levenberg J *et al* (2016) TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv*:1603.04467

Agathocleous M, Christodoulou G, Promponas V, Christodoulou C, Vassiliades V, Antoniou A (2010) Protein secondary structure prediction with bidirectional recurrent neural nets: can weight updating for each residue enhance performance? In *Artificial Intelligence Applications and Innovations*, Papadopoulos H, Andreou AS, Bramer M (eds), Vol. 339, pp 128 – 137. Berlin Heidelberg: Springer

Alain G, Bengio Y, Rifai S (2012) Regularized auto-encoders estimate local statistics. In *Proc. CoRR*, pp 1 – 17

Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. *Nature* 506: 494 – 497

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33: 831 – 838

Angermueller C, Lee H, Reik W, Stegle O (2016) Accurate prediction of single-cell DNA methylation states using deep learning. *bioRxiv*. doi: 10.1101/055715

Asgari E, Mofrad MRK (2015) ProtVec: a continuous distributed representation of biological sequences. *PLoS ONE* 10: e0141287

Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10: e0130140

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. *arXiv*:1409.0473

Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I, Bergeron A, Bouchard N, Warde-Farley D, Bengio Y (2012) Theano: new features and speed improvements. *arXiv*:1211.5590

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y (2015) Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 347: 664 – 667

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12: R10

Bengio Y, Lamblin P, Popovici D, Larochelle H (2007) Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, Schölkopf B, Platt B, Hofmann T (eds), Vol. 19, pp 153. Cambridge: MIT Press

Bengio Y (2012) Practical recommendations for gradient-based training of deep architectures. In *Neural networks: tricks of the trade*, Montavon G, Orr G, Müller K-R (eds), pp 437 – 478. Berlin Heidelberg: Springer

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. *Pattern Anal Mach Intell IEEE Trans* 35: 1798 – 1828

Bergstra J, Cox DD (2013) Hyperparameter optimization and boosting for classifying facial expressions: how good can a "Null" model be? *arXiv*:1306.3476

Che Z, Purushotham S, Khemani R, Liu Y (2015) Distilling knowledge from deep networks with applications to healthcare domain. *arXiv*:1512.03542

Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 12: R15

Ciresan D, Giusti A, Gambardella LM, Schmidhuber J (2012) Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pp 2843 – 2851. Cambridge: MIT Press

Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J (2013) Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp 411 – 418. Berlin Heidelberg: Springer

Collobert R, Kavukcuoglu K, Farabet C (2011) Torch7: a matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*

Dahl GE, Jaitly N, Salakhutdinov R (2014) Multi-task neural networks for QSAR predictions. *arXiv*:1406.1231

Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pp 2933 – 2941. Cambridge: MIT Press

Deng L (2014) Deep learning: methods and applications. *Found Trends® Signal Process* 7: 197 – 387

Deng L, Togneri R (2015) Deep dynamic models for learning hidden representations of speech features. In *Speech and audio processing for coding, enhancement and recognition*, Ogunfunmi T, Togneri R, Narasimha M (eds), pp 153 – 195. New York: Springer

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2013) Decaf: a deep convolutional activation feature for generic visual recognition. *arXiv*:1310.1531

Eduati F, Mangravite LM, Wang T, Tang H, Bare JC, Huang R, Norman T, Kellen M, Menden MP, Yang J, Zhan X, Zhong R, Xiao G, Xia M, Abdo N, Kosyk O, Collaboration N-N-UDT, Friend S, Dearry A, Simeonov A *et al* (2015) Prediction of human population responses to toxic compounds by a collaborative competition. *Nat Biotechnol* 33: 933 – 940

Eickholt J, Cheng J (2012) Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics* 28: 3066 – 3072

Eickholt J, Cheng J (2013) DNdisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics* 14: 88

Fakoor R, Ladhak F, Nazi A, Huber M (2013) Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare*, Atlanta, GA: JMLR, W&CP

Farley B, Clark W (1954) Simulation of self-organizing systems by digital computer. *Trans IRE Profess Group Inf Theory* 4: 76 – 84

Ferrari A, Lombardi S, Signoroni A (2015) Bacterial colony counting by Convolutional Neural Networks. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp 7458 – 7461. New York: IEEE

Gawehn E, Hiss JA, Schneider G (2016) Deep learning in drug discovery. *Mol Informatics* 35: 3 – 14

Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, Arepalli S, Dillman A, Rafferty IP, Troncoso J (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 6: e1000952

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580 – 587. New York: IEEE

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp 249 – 256. JMLR Conference Proceedings

© 2016 The Authors

Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp 315 – 323. JMLR Conference Proceedings

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. Cambridge: MIT Press in preparation

Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 6645 – 6649. New York: IEEE

Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A, Heidari N, Euskirchen G, Huber W, Pritchard JK, Bustamante CD, Steinmetz LM, Kundaje A, Snyder M (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 162: 1051 – 1065

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. *Math Intell* 27: 83 – 85

He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. *arXiv*:1512.03385

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313: 504 – 507

Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18: 1527 – 1554

Hinton GE (2012) A practical guide to training restricted Boltzmann machines. In *Neural networks: tricks of the trade*, Montavon G, Orr G, Müller K-R (eds), pp 599 – 619. Heidelberg Berlin: Springer

Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Process Mag IEEE* 29: 82 – 97

Hubel D, Wiesel T (1963) Shape and arrangement of columns in cat's striate cortex. *J Physiol* 165: 559

Hubel DH, Wiesel TN (1970) The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J Physiol* 206: 419

Hutter F, Hoos HH, Leyton-Brown K (2011) Sequential model-based optimization for general algorithm configuration. In *Learning and intelligent optimization*, Coello Coello CA (ed.), Vol. 6683, pp 507 – 523. Heidelberg Berlin: Springer

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv*:1502.03167

Jain V, Murray JF, Roth F, Turaga S, Zhigulin V, Briggman KL, Helmstaedter MN, Denk W, Seung HS (2007) Supervised learning of image restoration with convolutional networks. In *2007 IEEE 11th International Conference on Computer Vision*, pp 1 – 8. New York: IEEE

Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y (2009) What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pp 2146 – 2153. New York: IEEE

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pp 675 – 678. New York: ACM

Kang HM, Ye C, Eskin E (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180: 1909 – 1925

Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M (2010) Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA* 107: 2926 – 2931

Kell DB (2005) Metabolomics, machine learning and modelling: towards an understanding of the language of cells. *Biochem Soc Trans* 33: 520 – 524

Kelley DR, Snoek J, Rinn J (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* doi:10.1101/gr.200535.115

Kingma DP, Welling M (2013) Auto-encoding variational bayes. *arXiv*:1312.6114

Kingma D, Ba J (2014) Adam: a method for stochastic optimization. *arXiv*:1412.6980

Koh PW, Pierson E, Kundaje A (2016) Denoising genome-wide histone ChIP-seq with convolutional neural networks. *bioRxiv* doi:10.1101/052118

Kraus OZ, Ba LJ, Frey B (2015) Classifying and segmenting microscopy images using convolutional multiple instance learning. *arXiv*:1511.05286v1

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp 1097 – 1105. Cambridge: MIT Press

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1: 541 – 551

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521: 436 – 444

Lee B, Lee T, Na B, Yoon S (2015) DNA-level splice junction prediction using deep recurrent neural networks. *arXiv*:1512.05135

Leung MKK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics* 30: i121 – i129

Leung MKK, Delong A, Alipanahi B, Frey BJ (2016) Machine learning in genomic medicine: a review of computational problems and data sets. In *Proceedings of the IEEE*, Vol. 104, pp 176 – 197. New York: IEEE

Li SZ (2009) *Markov random field modeling in image analysis*. Berlin Heidelberg: Springer Science & Business Media

Li J, Ching T, Huang S, Garmire LX (2015) Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics* 16(Suppl. 5): S10

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. *Nat Rev Genet* 16: 321 – 332

Lipton ZC (2015) A critical review of recurrent neural networks for sequence learning. *arXiv*:1506.00019

Lipton ZC, Kale DC, Elkan C, Wetzell R (2015) Learning to diagnose with LSTM recurrent neural networks. *arXiv*:1511.03677

Lyons J, Dehzangi A, Heffernan R, Sharma A, Paliwal K, Sattar A, Zhou Y, Yang Y (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J Comput Chem* 35: 2040 – 2046

Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. *Mol Pharm* 13: 1445 – 1454

Märtens K, Hallin J, Warringer J, Liti G, Parts L (2016) Predicting quantitative traits from genome and phenome with near perfect accuracy. *Nat Commun* 7: 11512

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5: 115 – 133

Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* 8: e61318

Michalski RS, Carbonell JG, Mitchell TM (2013) *Machine learning: an artificial intelligence approach*. Berlin Heidelberg: Springer Science & Business Media

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464: 773 – 777

 © 2016 The Authors

Murphy KP (2012) *Machine learning: a probabilistic perspective*. Cambridge: MIT Press

Nesterov Y (1983) A method of solving a convex programming problem with convergence rate O (1/k2). *Soviet Math Doklady* 27: 372 – 376

Nesterov Y (2013) *Introductory lectures on convex optimization: a basic course*, Vol. 87, Berlin Heidelberg: Springer Science & Business Media

Ning F, Delhomme D, LeCun Y, Piano F, Bottou L, Barbano PE (2005) Toward automatic phenotyping of developing embryos from videos. *Image Process IEEE Trans* 14: 1360 – 1371

Park Y, Kellis M (2015) Deep learning for regulatory genomics. *Nat Biotechnol* 33: 825 – 826

Pärnamaa T, Parts L (2016) Accurate classification of protein subcellular localization from high throughput microscopy images using deep learning. *bioRxiv* doi:10.1101/050757

Parts L, Stegle O, Winn J, Durbin R (2011) Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet* 7: e1001276

Parts L, Liu YC, Tekkedil MM, Steinmetz LM, Caudy AA, Fraser AG, Boone C, Andrews BJ, Rosebrock AP (2014) Heritability and genetic basis of protein level variation in an outbred population. *Genome Res* 24: 1363 – 1370

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768 – 772

Rakitsch B, Stegle O (2016) Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biol* 17: 33

Rampasek L, Goldenberg A (2016) TensorFlow: biology's gateway to deep learning? *Cell Syst* 2: 12 – 14

Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp 806 – 813. New York: IEEE

Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp 234 – 241. Heidelberg Berlin: Springer

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65: 386

Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. *Cogn Model* 5: 1

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115: 211 – 252

Salakhutdinov R, Larochelle H (2010) Efficient learning of deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pp 693 – 700. JMLR Conference Proceedings

Salakhutdinov R, Hinton G (2012) An efficient learning procedure for deep Boltzmann machines. *Neural Comput* 24: 1967 – 2006

Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61: 85 – 117

Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv*:1312.6034

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv*:1409.1556

Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp 2951 – 2959. Cambridge: MIT Press

Sønderby SK, Winther O (2014) Protein secondary structure prediction with long short term memory networks. *arXiv*:1412.7828

Spencer M, Eickholt J, Cheng J (2015) A deep learning network approach to *ab initio* protein secondary structure prediction. *IEEE/ACM Trans Comput Biol Bioinformatics* 12: 103 – 112

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: the all convolutional net. *arXiv*:1412.6806

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mac Learn Res* 15: 1929 – 1958

Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 6: e1000770

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* 10: 2997 – 3011

Sutskever I (2013) Training recurrent neural networks. PhD Thesis, Graduate Department of Computer Science, University of Toronto, Toronto, Canada

Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp 1139 – 1147. JMLR Conference Proceedings

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp 3104 – 3112. Cambridge: MIT Press

Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J (2013) Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* 17: 595 – 610

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015a) Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 1 – 9. New York: IEEE

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015b) Rethinking the inception architecture for computer vision. *arXiv*:1512.00567

Team TTD, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A (2016) Theano: a python framework for fast computation of mathematical expressions. *arXiv*:1605.02688

Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mac Learn Res* 11: 3371 – 3408

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 3156 – 3164. New York: IEEE

Wang K, Cao K, Hannenhalli S (2015) Chromatin and genomic determinants of alternative splicing. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pp 345 – 354. New York: ACM

Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A, Romano-Palumbo L, Planchon A, Bielser D, Padioleau I, Udin G, Thurnheer S, Hacker D, Hernandez N, Reymond A, Deplancke B *et al* (2015) Population variation and genetic control of modular chromatin architecture in humans. *Cell* 162: 1039 – 1050

Xie Y, Xing F, Kong X, Su H, Yang L (2015) Beyond classification: structured regression for robust cell detection using convolutional neural network. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp 358 – 365. Heidelberg Berlin: Springer

© 2016 The Authors

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347: 1254806

Xiong C, Merity S, Socher R (2016) Dynamic memory networks for visual and textual question answering. *arXiv*:1603.01417

Xu R, Wunsch D II, Frank R (2007) Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Trans Comput Biol Bioinformatics* 4: 681 – 692

Xu Y, Mo T, Feng Q, Zhong P, Lai M, Chang EI (2014) Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp 1626 – 1630. New York: IEEE

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. *arXiv*:1502.03044

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In *Advances in neural information processing systems 27*, Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds), pp 3320 – 3328. Cambridge: MIT Press

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In Computer vision–ECCV 2014, pp 818 – 833. Heidelberg Berlin: Springer

Zhang W, Li R, Zeng T, Sun Q, Kumar S, Ye J, Ji S (2015) Deep model based transfer and multi-task learning for biological image analysis. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 1475 – 1484. New York: ACM

Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12: 931 – 934

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Ten quick tips for deep learning in biology

Benjamin D. Lee[1,2,3]*, **Anthony Gitter**[4,5], **Casey S. Greene**[6,7,8], **Sebastian Raschka**[9], **Finlay Maguire**[10], **Alexander J. Titus**[11,12], **Michael D. Kessler**[13,14], **Alexandra J. Lee**[6,15], **Marc G. Chevrette**[16], **Paul Allen Stewart**[17], **Thiago Britto-Borges**[18,19], **Evan M. Cofer**[20,21], **Kun-Hsing Yu**[22,23], **Juan Jose Carmona**[24], **Elana J. Fertig**[13,25], **Alexandr A. Kalinin**[26,27], **Brandon Signal**[28], **Benjamin J. Lengerich**[29], **Timothy J. Triche, Jr.**[30,31,32], **Simina M. Boca**[33,34,35,36]

1 In-Q-Tel Labs, Arlington, Virginia, United States of America, **2** School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **4** Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **5** Morgridge Institute for Research, Madison, Wisconsin, United States of America, **6** Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **7** Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado, United States of America, **8** Center for Health AI, University of Colorado School of Medicine, Aurora, Colorado, United States of America, **9** Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **10** Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada, **11** University of New Hampshire, Manchester, New Hampshire, United States of America, **12** Bioeconomy. XYZ, Manchester, New Hampshire, United States of America, **13** Department of Oncology, Johns Hopkins University, Baltimore, Maryland, United States of America, **14** Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America, **15** Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **16** Wisconsin Institute for Discovery and Department of Plant Pathology, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **17** Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, United States of America, **18** Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology, University Hospital Heidelberg, Heidelberg, Germany, **19** Department of Internal Medicine III (Cardiology, Angiology, and Pneumology), University Hospital Heidelberg, Heidelberg, Germany, **20** Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America, **21** Graduate Program in Quantitative and Computational Biology, Princeton University, Princeton, New Jersey, United States of America, **22** Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, United States of America, **23** Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **24** Philips Healthcare, Cambridge, Massachusetts, United States of America, **25** Department of Biomedical Engineering, Department of Applied Mathematics and Statistics, Convergence Institute, Johns Hopkins University, Baltimore, Maryland, United States of America, **26** Medical Big Data Group, Shenzhen Research Institute of Big Data, Shenzhen, China, **27** Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America, **28** School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Australia, **29** Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America, **30** Center for Epigenetics, Van Andel Research Institute, Grand Rapids, Michigan, United States of America, **31** Department of Pediatrics, College of Human Medicine, Michigan State University, East Lansing, Michigan, United States of America, **32** Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America, **33** Innovation Center for Biomedical Informatics, Georgetown University Medical Center, District of Columbia, United States of America, **34** Department of Oncology, Georgetown University Medical Center, Washington, DC, United States of America, **35** Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, DC, United States of America, **36** Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, Washington, DC, United States of America

* benjamindlee@me.com

Check for updates

OPEN ACCESS

Citation: Lee BD, Gitter A, Greene CS, Raschka S, Maguire F, Titus AJ, et al. (2022) Ten quick tips for deep learning in biology. PLoS Comput Biol 18(3): e1009803. https://doi.org/10.1371/journal.pcbi.1009803

Editor: Francis Ouellette, McGill University, CANADA

Published: March 24, 2022

Copyright: © 2022 Lee et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: A.G. was funded by the National Science Foundation (DBI 1553206) and National Institutes of Health (R01GM135631). C.S.G. was funded by the National Institutes of Health (R01 HG010067) and the Gordon and Betty Moore Foundation (GBMF 4552). S.R. was funded by the Wisconsin Alumni Foundation (AAD5912). F.M. was funded by the Donald Hill Family Fellowship. M.D.K. was funded by the National Institutes of Health (R01DE027809). A.J.L. was funded by the Gordon and Betty Moore Foundation (GBMF 4552). M.G.C. was funded by Grant 2020-67012-31772 (accession 1022881) from the USDA National Institute of Food and Agriculture. P.A.S. was funded by the Moffitt Cancer Center Support Grant (P30-CA076292). E.M.C. was funded by the

Introduction

Machine learning is a modern approach to problem-solving and task automation. In particular, machine learning is concerned with the development and applications of algorithms that

National Institutes of Health (T32 HG003284) and the National Science Foundation Graduate Research Fellowship Program. K.H.Y. was funded by the Blavatnik Center for Computational Biomedicine Award and the National Institute of General Medical Sciences (R35GM142879). E.J.F. was funded by the Lustgarten Foundation, Allegheny Health Network, Emerson Foundation (640183), National Cancer Institute (U01CA212007, U01CA253403, P30CA006973); and National Institute of Dental and Craniofacial Research (R01DE027809). T.J.T. was funded by the National Institute of Allergy and Infectious Diseases (R21AI153997), Michelle Lunn Hope Foundation, Folz Fund for Cancer Research, and Grand Rapids Community Foundation. S.M.B. was funded by the National Institutes of Health (R21 CA220398). The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and have the following conflicts: AG filed a patent application with the Wisconsin Alumni Research Foundation related to classifying activated T cells. KHY is the inventor of a quantitative pathology analytical system (U.S. Patent 10,832,406). This invention is not related to this work. EJF is a Scientific Advisory Board member at Viosera Therapeutics. AAK is a co-inventor on 4 patent applications related to machine learning applications in biology.

can recognize patterns in data and use them for predictive modeling, as opposed to having domain experts developing rules for prediction tasks manually. Artificial neural networks are a particular class of machine learning algorithms and models that evolved into what is now described as "deep learning". Deep learning encompasses neural networks with many layers and the algorithms that make them perform well. These neural networks comprise artificial neurons arranged into layers and are modeled after the human brain, even though the building blocks and learning algorithms may differ [1]. Each layer receives input from previous layers (the first of which represents the input data), and then transmits a transformed version of its own weighted output that serves as input into subsequent layers of the network. Thus, the process of "training" a neural network is the tuning of the layers' weights to minimize a cost or loss function that serves as a surrogate of the prediction error. The loss function is differentiable so that the weights can be automatically updated to attempt to reduce the loss. Deep learning uses artificial neural networks with many layers (hence the term "deep"). Given the computational advances made in the last decade, it can now be applied to massive data sets and in innumerable contexts. In many circumstances, deep learning can learn more complex relationships and make more accurate predictions than other methods. Therefore, deep learning has become its own subfield of machine learning. In the context of biological research, it has been increasingly used to derive novel insights from high-dimensional biological data [2]. For example, deep learning has been used to predict protein–drug binding kinetics [3], to identify the lab-of-origin of synthetic DNA [4], and to uncover the facial phenotypes of genetic disorders [5].

To make the biological applications of deep learning more accessible to scientists who have some experience with machine learning, we solicited input from a community of researchers with varied biological and deep learning interests. These individuals collaboratively contributed to this manuscript's writing using the GitHub version control platform [6] and the Manubot manuscript generation toolset [7]. The goal was to articulate a practical, accessible, and concise set of guidelines and suggestions to follow when using deep learning (Fig 1). For readers who are new to machine learning, we recommend reviewing general machine learning principles [8] before getting started with deep learning.

In the course of our discussions, several themes became clear: the importance of understanding and applying machine learning fundamentals as a baseline for utilizing deep learning, the necessity for extensive model comparisons with careful evaluation, and the need for critical thought in interpreting results generated by deep learning, among others. The major similarities between deep learning and traditional computational methods also became apparent. Although deep learning is a distinct subfield of machine learning, it is still a subfield. It is subject to the many limitations inherent to machine learning, and most best practices for machine learning [9–11] also apply to deep learning. As with all computational methods, deep learning should be applied in a systematic manner that is reproducible and rigorously tested. Ultimately, the tips we collate range from high-level guidance to best practices for implementation. It is our hope that they will provide actionable, deep learning–specific instructions for both new and experienced deep learning practitioners. By making deep learning more accessible for use in biological research, we aim to improve the overall usage and reporting quality of deep learning in the literature and to enable increasing numbers of researchers to utilize these state-of-the art techniques effectively and accurately.

Tip 1: Decide whether deep learning is appropriate for your problem

In recent years, the number of projects and publications implementing deep learning in biology has risen tremendously [12–14]. This trend is likely driven by deep learning's usefulness

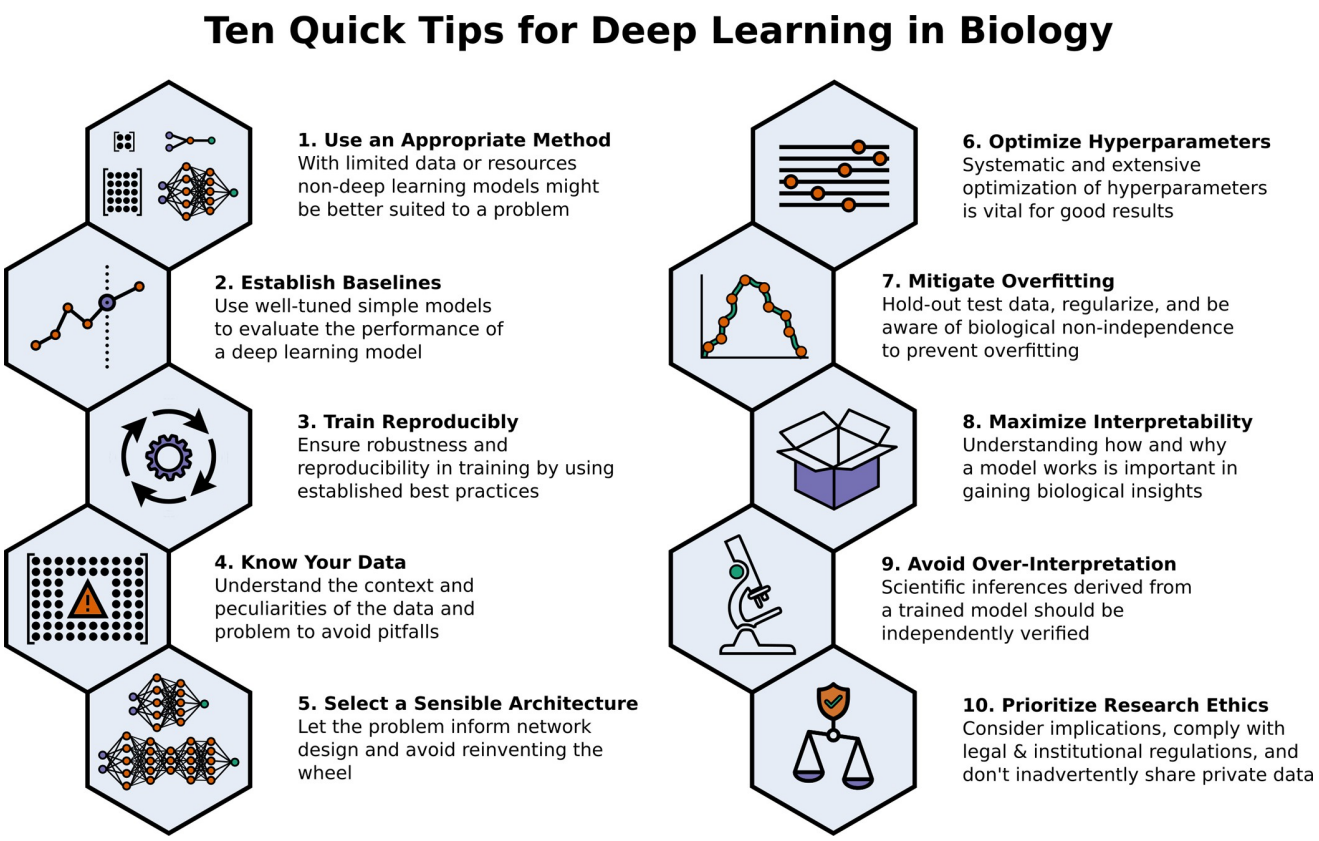


Fig 1. A summary overview of the 10 tips for using deep learning in biological research.

https://doi.org/10.1371/journal.pcbi.1009803.g001

across a range of scientific questions and data modalities and can contribute to the appearance of deep learning as a panacea for nearly all modeling problems. Indeed, neural networks are universal function approximators and derive tremendous power from this theoretical capacity to learn any function [15,16]. However, in reality, deep learning is not suited to every modeling situation and can be significantly limited by its large demands for data, computing power, programming skill, and modeling expertise.

While large amounts of high-quality data may be available in the areas of biology where data collection is thoroughly automated, such as DNA sequencing, areas of biology that rely on manual data collection may not possess enough data to train and apply deep learning models effectively. Methods that try to increase the amount of training data, such as data augmentation (in which existing data is slightly manipulated in an attempt to yield "new" samples) [17] and weak supervision (in which simple labeling heuristics are combined to produce noisy, probabilistic labels) [18] are valuable for small- to medium-scale datasets. For example, when classifying microalgae using models trained on 21,000 images, data augmentation improved the prediction accuracy by 17% [19]. In a text detection context based on a small dataset of 229 fully annotated scene text images, weakly supervised learning improved the precision by 11% [20]. However, methods cannot overcome substantial data shortages in many practical scenarios, and recent research investigating machine learning methods in neuroimaging studies of depression suggests that high prediction accuracies obtained from small datasets may be caused by misestimation due to insufficient test dataset sizes [21].

In the fields of computer vision and natural language processing, deep neural networks are routinely trained on sample sizes ranging from hundreds of thousands to millions of training examples [22,23]. Datasets of this size are often not available in many biological contexts. Still, it has been found that, under certain circumstances, deep learning can be considered for datasets with only 100 samples per class [24]. Nonetheless, deep learning is generally best suited for datasets that contain orders of magnitude more samples.

Training deep learning models often requires extensive computing infrastructure and patience to achieve state-of-the-art performance [25]. In some deep learning contexts, such as generating human-like text, state-of-the-art models have over 100 billion parameters [26] and require very costly and time-consuming training procedures [27]. These types of large language models are being used in biology to learn representations of protein sequences [28–30]. Even though most deep learning applications in biology rarely require this much training, they can still require computational resources beyond those available on consumer-grade devices such as laptops or office desktops. Specialized hardware such as discrete graphics processing units (GPUs) and custom deep learning accelerators can dramatically reduce the time and cost required to train models [14]. Still, this hardware is not universally accessible, and cloud-based rentals add additional cost and complexity. These specialized hardware solutions are likely to be more broadly available as deep learning becomes more popular. For example, recent-generation iPhones already have such hardware. In contrast to the large-scale computational demands of deep learning, traditional machine learning models can often be trained on laptops (or even on a $5 computer [31]) in seconds to minutes. Therefore, due to this enormous disparity in resource demand alone, traditional machine learning approaches may be desirable in various biological applications.

Beyond requiring more data and computational capacity, building and training deep learning models often requires more expertise than training traditional machine learning models. While popular programming frameworks for deep learning such as TensorFlow [32] and PyTorch [33] allow users to create and deploy entirely novel model architectures, this flexibility combined with the rapid development of the deep learning field has resulted in large and complex frameworks that can be daunting to new users. For readers new to software development but experienced in biology, gaining computational skills while also interfacing with such complex industrial-grade tools can be a prohibitive challenge. Conversely, traditional machine learning methods are generally more straightforward to apply and are also more accessible through popular frameworks [34]. Furthermore, there are currently more tools for automating the model selection and training process for traditional machine learning models than for deep learning models. For example, automated machine learning (Auto ML) tools, such as TPOT [35] and Turi Create [36], are able to test and optimize multiple machine learning models automatically and can allow users to achieve competitive performance with only a few lines of code. There are efforts underway to extend these and other automation frameworks to reduce the expertise required to build and use deep learning models. TPOT, Turi Create, and Auto Keras [37] are already capable of abstracting away much of the programming required for "standard" deep learning tasks, and high-level interfaces such as Keras [38] and Fastai [39] make it increasingly straightforward to design and test custom deep learning architectures. In the future, projects such as these are likely to make deep learning accessible to a much wider range of researchers.

Despite these limitations, deep learning is strongly indicated over traditional machine learning for specific research questions and problems. In general, these include problems that feature hidden patterns across the data, complex relationships, and interrelated variables. Problems in computer vision and natural language processing often exhibit these very features, which helps explain why these areas were some of the first to experience significant breakthroughs during the recent deep learning revolution [40]. As long as large amounts of accurate

and labeled data are available, applications to areas of life sciences with related data characteristics, such as genetic medicine [41], radiology [42], microscopy [43], and pharmacovigilance [44], are similarly likely to benefit from deep learning techniques. For example, Ferreira and colleagues used deep learning to recognize individual birds from images [45] despite this problem being very difficult historically. By combining automatic data collection using radio-frequency identification tags with data augmentation and transfer learning, the authors were able to use deep learning to achieve 90% accuracy across several species. Another research area where deep learning excels is generative modeling, where new samples are created based on the training data [46]. For example, deep learning can generate realistic compendia of gene expression samples [47]. One other area of machine learning that has been revolutionized by deep learning is reinforcement learning, which is concerned with training agents to interact with an environment [48]. Reinforcement learning has been applied to design drug like small molecules [49]. Overall, initial evaluation as to whether similar problems (including analogous ones in other domains) have been solved successfully using deep learning can inform researchers about the potential for deep learning to address their needs.

On the other hand, depending on the amount and type of data available and the nature of the problem set, deep learning may not always outperform conventional methods [50,51]. As an illustration, Rajkomar and colleagues [52] found that simpler baseline models achieved performance comparable with deep learning in several clinical prediction tasks using electronic health records. Another example is provided by Koutsoukas and colleagues, who benchmarked several traditional machine learning approaches against deep neural networks for modeling bioactivity data on moderately sized datasets [53]. The researchers found that while well-tuned deep learning approaches generally tend to outperform conventional classifiers, simple methods such as Naive Bayes classification tend to outperform deep learning as the dataset's noise increases. Similarly, Chen and colleagues [54] tested deep learning and a variety of traditional machine learning methods such as logistic regression and random forests on 5 different clinical datasets. They found that traditional methods matched or exceeded the accuracy of the deep learning model in all cases despite requiring an order of magnitude less training time.

In conclusion, deep learning should only be used after a robust consideration of its strengths and weaknesses for the problem at hand. After choosing deep learning as a potential solution, practitioners should still consider traditional methods as performance baselines.

Tip 2: Use traditional methods to establish performance baselines

Deep learning requires practitioners to consider a larger number and variety of tuning parameters (that is, algorithmic settings) than more traditional machine learning methods. These settings are often called hyperparameters. Their extensiveness can make it easy to fall into the trap of performing an unnecessarily convoluted analysis. Hence, before applying deep learning to a given problem, it is ideal to implement simpler models with fewer hyperparameters at the beginning of each study [11]. Such models include logistic regression, random forests, k-nearest neighbors, Naive Bayes, and support vector machines. They can help to establish baseline performance expectations and the difficultly of a specific prediction problem. While performance baselines available from existing literature can also serve as helpful guides, an implementation of a simpler model that uses the same software framework as planned for deep learning can greatly help with assessing the correctness of data processing steps, performance evaluation pipelines, resource requirement estimates, and computational performance estimates. Furthermore, in some cases, it can even be useful to combine simpler baseline models with deep neural networks, as such hybrid models can improve generalization performance, model interpretability, and confidence estimation [55,56].

Another potential pitfall arises from comparing the performance of baseline conventional models trained with default settings with the performance of deep learning models that have undergone rigorous tuning and optimization. Since conventional off-the-shelf machine learning algorithms (for example, support vector machines and random forests) are also likely to benefit from hyperparameter tuning, such incongruity prevents the comparison of equally optimized models and can lead to false conclusions about model efficacy. Hu and Greene [57] discuss this under the umbrella of what they call the "Continental Breakfast Included" effect. They describe how the unequal tuning of hyperparameters across different learning algorithms can especially skew evaluation when the performance of an algorithm varies substantially with modest changes to its hyperparameters. Therefore, practitioners should tune the settings of both traditional machine learning and deep learning–based methods before making claims about relative performance differences. Performance comparisons among machine learning and deep learning models are only informative when the models are equally well optimized.

In sum, practitioners are encouraged to create and fully tune several traditional models and standard pipelines before implementing a deep learning model.

Tip 3: Understand the complexities of training deep neural networks

Correctly training deep neural networks is nontrivial. There are many different options and potential pitfalls at every stage. To get good results, one must often train networks across a wide range of different hyperparameter settings. Such training can be made more difficult by the demanding nature of these deep networks, which often require extensive time investments into tuning and computing infrastructure to achieve state-of-the-art performance [25]. Furthermore, this experimentation is often noisy, which necessitates increased repetition and exacerbates the challenges inherent to deep learning. On the whole, all code, random seeds, parameters, and results must be carefully corralled using general coding standards and best practices (for example, version control [58] and continuous integration [59]) to be reproducible and interpretable [60–62]. For application-based research, this organization is also fundamental to the efficient sharing of research work and the ability to keep models up to date as new data becomes available.

The use of nondeterministic algorithms, often by default, is a specific reproducibility pitfall that is often missed in applying deep learning. For example, GPU acceleration libraries like CUDA/CuDNN, which facilitate the parallelized computing powering state-of-the-art deep learning, often use default algorithms that produce different outcomes from iteration to iteration even with the same hardware and software. Achieving reproducibility in this context requires explicitly specifying the use of deterministic algorithms, which are typically available within deep learning libraries [63]. This step is distinct from and in addition to the setting of random seeds that typically achieve reproducibility by controlling pseudorandom deterministic procedures such as shuffling and initialization.

Similar to the suggestions above about starting with simpler models, starting with relatively small networks and then increasing the size and complexity as needed can help prevent practitioners from wasting significant time and resources on running highly complex models that feature numerous unresolved problems. Again, practitioners must be aware of the choices made implicitly (that is, by default settings) by deep learning libraries. These seemingly trivial details, such as the selection of optimization algorithm, can have significant effects on model performance. For example, adaptive methods often lead to faster convergence during training but may lead to worse generalization performance on independent datasets [64]. These nuanced elements are easy to overlook, but it is critical to consider them carefully and to evaluate their potential impact.

In short, researchers should use smaller and simpler networks to enable faster prototyping, follow general software development best practices to maximize reproducibility, and check software documentation to understand default choices.

Tip 4: Know your data and your question

Having a well-defined scientific question and a clear analysis plan is crucial for carrying out a successful deep learning project. Just like it would be inadvisable to set foot in a laboratory and begin experiments without having a defined endpoint, a deep learning project should not be undertaken without defined goals. Foremost, it is important to assess if a dataset exists that can answer the biological question of interest using a deep learning–based approach. If so, obtaining this data (and associated metadata) and reviewing the study protocol should be pursued as early on in the project as possible. This can help to ensure that data is as expected and can prevent the wasted time and effort that occur when issues are discovered later on in the analytic process. For example, a publication or resource might purportedly offer an appropriate dataset that is found to be inadequate upon acquisition. The data may be unstructured when it is supposed to be structured, crucial metadata such as sample stratification might be missing, or the usable sample size may be different than expected. Any of these data issues might limit a researcher's ability to use deep learning to address the biological question at hand or might otherwise require adjustment before it can be used. Data collection should also be carefully documented, or a data collection protocol should be created and specified in the project documentation.

Information about the resources used, download dates, and dataset versions are critical to preserve. Doing so will help to minimize operational confusion and will increase the reproducibility of the analysis. Best practices for reproducibility also include sharing the collected dataset and metadata upon publication of the study, ideally in a public dataset repository if there are no ethical or privacy concerns and no copyright restrictions. While recommended and recognized dataset repositories may differ across disciplines, a list of general dataset repositories includes the Dryad repository [65] (https://datadryad.org/), Fig share [66] (https://figshare.com), Zenodo [67] (https://zenodo.org), and the Open Science Framework [68] (https://osf.io). In addition, Gundersen and colleagues [69] provide useful checklists summarizing best data sharing practices for reproducible research and open science.

Once the dataset is obtained, it is important to learn why and how the data was collected before beginning the analysis. The standardized metadata that exists in many fields can help with this (for example, see [70]). If at all possible, consulting with a subject matter expert who has experience with the type of data being used will minimize guesswork and likely increase the success rate of a deep learning project. For example, one might presume that data collected to test the impact of an intervention are derived from a randomized controlled trial. However, this is not always the case, as ethical or practical concerns often necessitate an observational study design that features prospectively or retrospectively collected data. In order to ensure similar distributions of important characteristics across study groups in the absence of randomization, such a study may have selected individuals in a fashion that best matches attributes such as age, gender, or weight. Passively collected datasets can have their own peculiarities, and other study designs can include samples that originate from the same study site, the oversampling of ethnic groups or zip codes, or sample processing differences. Such information is critical to accurate data analysis, and so it is imperative that practitioners learn about study design assumptions and data specificities prior to performing modeling. Other study design considerations that should not be overlooked include knowing whether a study involves biological or technical replicates or both. For example, the existence in a dataset of samples collected from the same individuals at different time points can have significant effects on analyses that make assumptions

about sample size and independence (that is, non independence can lower the effective sample size). Another potential issue is the existence of systematic biases, which can be induced by confounding variables and can lead to artifacts or so-called "batch effects." Consequently, models may learn to rely on the correlations that these systematic biases underpin, even though they are irrelevant to the scientific context of the study. This can lead to misguided predictions and misleading conclusions [71]. Unsupervised learning and other exploratory analyses can help identify such biases in these datasets before applying a deep learning model.

Overall, practitioners should thoroughly study their data and understand its context and peculiarities before moving on to performing deep learning.

Tip 5: Choose an appropriate data representation and neural network architecture

Neural network architecture refers to the number and types of layers in the network and how they are connected. While certain best practices have been established by the research community [72], architecture design choices remain largely problem-specific and are vastly empirical efforts requiring extensive experimentation. Furthermore, as deep learning is a quickly evolving field, many recommendations are often short-lived and are frequently replaced by newer insights supported by recent empirical results. This is further complicated by the fact that many recommendations do not generalize well across different problems and datasets. Therefore, choosing how to represent data and design an architecture is closer to an art than a science. That said, there are some general principles to follow when experimenting.

First and foremost, knowledge of the available data and scientific question should inform data representation and architectural design choices. For example, if the dataset is an array of measurements with no natural ordering of inputs (such as gene expression data), multilayer perceptrons may be effective. These are the most basic type of neural network. They are able to learn complex nonlinear relationships across the input data despite their relative simplicity. Similarly, if the dataset is comprised of images, convolutional neural networks (CNNs) are a good choice because they emphasize local structures and adjacency within the data. CNNs may also be a good choice for learning on sequences. Recent empirical evidence suggests that they can outperform canonical sequence learning techniques such as recurrent neural networks and the closely related long short-term memory networks in some cases [73]. Transformers are powerful sequence models [74] but require extensive data and computing power to train from scratch. Accessible high-level overviews of different neural network architectures are provided in [75,76].

Deep learning models typically benefit from increasing the amount of labeled training data. Large amounts of data help to avoid overfitting and increase the likelihood of achieving top performance on a given task. If there is not enough data available to train a well-performing model, transfer learning should be considered. In transfer learning, a model whose weights were generated by training on another dataset is used as the starting point for training [77]. Transfer learning is most useful when the pre training and target datasets are of similar nature [77]. For this reason, it is important to search for similar datasets that are already available. However, even when similar datasets do not exist, transferring features can still improve model performance compared with random feature initialization. For example, Rajkomar and colleagues showed advantages of ImageNet pre training [78] for a model that is applied to grayscale medical image classification [79]. Pre trained models can be obtained from public repositories, such as Kipoi [80] for genomics or Hugging Face [81] for biomedical text [82], protein sequences [29], and chemicals [83]. Moreover, learned features could be helpful even when a pre training task is different from a target task [84].

Recently, the concept of self-supervised learning, which is closely related to pre training and transfer learning, has seen an increase in popularity [85]. Self-supervised learning

leverages large amounts of unlabeled data and uses naturally available information as labels for supervised learning. Thus, self-supervised learning is sometimes also described as autonomous supervised learning. Using self-supervised learning, a model can be pre trained on a related task before it is trained on the target task. Another related approach is multitask learning, which simultaneously trains a network for multiple separate tasks that share features. In fact, multitask learning can be used separately or even in combination with transfer learning [86].

Practitioners should base the neural network's architecture on knowledge of the problem and take advantage of similar existing data or pre trained models.

Tip 6: Tune your hyperparameters extensively and systematically

Given at least one hidden layer, a nonlinear activation function, and a large number of hidden units, multilayer neural networks can approximate arbitrary continuous functions that relate input and output variables [16,87]. Deeper architectures that feature additional hidden layers and an increasing number of overall hidden units and learnable weight parameters (the so-called increasing "capacity" of neural networks) allow for solving increasingly complex problems. However, this increased capacity results in many more parameters to fit and hyperparameters to tune, which can pose additional challenges during model training. In general, one should expect to systematically evaluate the impact of numerous hyperparameters when applying deep neural networks to new data or challenges. Hyperparameters typically manifest as choices of optimization algorithms, loss function, learning rate, activation functions, number of hidden layers and hidden units, size of the training batches, and weight initialization schemes. Moreover, additional hyperparameters are introduced by common techniques that facilitate training deeper architectures. These include regularization penalties, dropout [88], and batch normalization [89], which can reduce the effect of the so-called vanishing or exploding gradient problem when working with deep neural networks.

This wide array of potential parameters can make it difficult to evaluate the extent to which neural network methods are well suited to solving a task. For example, it can be unclear to practitioners whether previously successful deep learning applications were the result of general model suitability for the data at hand or interactions between unique data attributes and specific hyperparameter settings. As a result, a lack of clarity about how extensive arrays of hyperparameters were tested or chosen can hamper method developers as they attempt to compare techniques. This effect also has implications for those seeking to use existing deep learning methods, as performance estimates from deep neural networks are often provided after tuning. The implication is that attaining performance numbers that match those reported in publications is likely to require significant effort toward temporally expensive hyperparameter optimization. Strategies for tuning hyperparameters include exhaustive grid search, random search, or Bayesian optimization and other specialized techniques. Tools such as Keras Tuner (https://keras-team.github.io/keras-tuner/) and Ray Tune (https://docs.ray.io/en/latest/tune/index.html) support best practices for hyperparameter optimization.

To get the best performance from your model, researchers should be sure to systematically optimize their hyperparameters on the training dataset and report both the selected hyperparameters and the hyperparameter optimization strategy.

Tip 7: Address deep neural networks' increased tendency to overfit the dataset

Overfitting is a challenge inherent to machine learning in general and is one of the most significant challenges you will face when applying deep learning specifically. Overfitting occurs when a model fits patterns in the training data so closely that it includes nongeneralizable

noise or nonscientifically relevant perturbations in the relationships it learns. In other words, the model fits patterns that are overly specific to the data it is training on rather than learning general relationships that hold across similar datasets. This subtle distinction is made clearer by seeing what happens when a model is tested on data to which it was not exposed during training: Just as a student who memorizes exam materials struggles to correctly answer questions for which they have not studied, a machine learning model that has overfit to its training data will perform poorly on unseen test data. Deep learning models are particularly susceptible to overfitting due to their relatively large number of parameters and associated representational capacity. Just as some students may have greater potential for memorization, deep learning models seem more prone to overfitting than machine learning models with fewer parameters. However, having a large number of parameters does not always imply that a neural network will overfit [90].

In general, one of the most effective ways to combat overfitting is to detect it in the first place. One way to do this is to split the main dataset being worked on into 3 independent parts: a training set, a tuning set (also commonly called a validation set in the machine learning literature), and a test set. These 3 partitions allow you to optimize models by iterating between model learning on the training set and hyperparameter evaluation on the tuning set without affecting the final model assessment on the test set. A researcher can compare the model's performance on the training and tuning data to assess how overfit (that is, nongeneralizable) the model is. The data used for testing should be "locked away" and used only once to evaluate the final model after all training and tuning steps are completed. This type of approach is necessary for evaluating the generalizability of models without the biases that can arise from learning and testing on the same data [91,92]. While a slight drop in performance from the training set to the test set is normal, a significant drop is a clear sign of overfitting. See Fig 2 for a visual demonstration of an overfit model that performs poorly on test data.

There are a variety of techniques to reduce overfitting, including data augmentation and regularization techniques [93,94]. Another way to reduce overfitting, as described by Chuang and Keiser, is to identify the baseline level of memorization that is occurring by training on data that has its labels randomly shuffled [95]. By comparing the model performance with the shuffled data to that achieved with the actual data [95], a practitioner can identify overfitting as a model that performs no better on real data. This suggests that any predictive capacity is not due to data-driven signal. One important caveat when working with partitioned data is the need to apply transformation and normalization procedures equally to all datasets. The parameters required for such procedures (for example, quantile normalization, a common standardization method when analyzing gene-expression data) should only be derived from the training data, and not from the tuning or test data.

Additionally, many conventional metrics for classification (for example, area under the receiver operating characteristic curve) have limited utility in cases of extreme class imbalance [96]. Consider an example of a dataset of mammograms in which 99% of the samples do not have breast cancer. A model could achieve 99% accuracy by classifying every sample as negative. Therefore, model performance should be evaluated with a carefully picked panel of relevant metrics that make minimal assumptions about the composition of the testing data [97]. One alternative approach is to use the precision-recall curve rather than the receiver operating characteristic since the former is more robust to class imbalance [98].

When working with biological and medical data, one must also carefully consider potential sources of bias and/or nonindependence when defining training and test sets. For example, a deep learning model for pneumonia detection in chest X-rays appeared to perform well within the hospitals providing the training data, but then failed to generalize to other hospitals [99]. This resulted from the deep learning model picking up on signals related to which hospital the

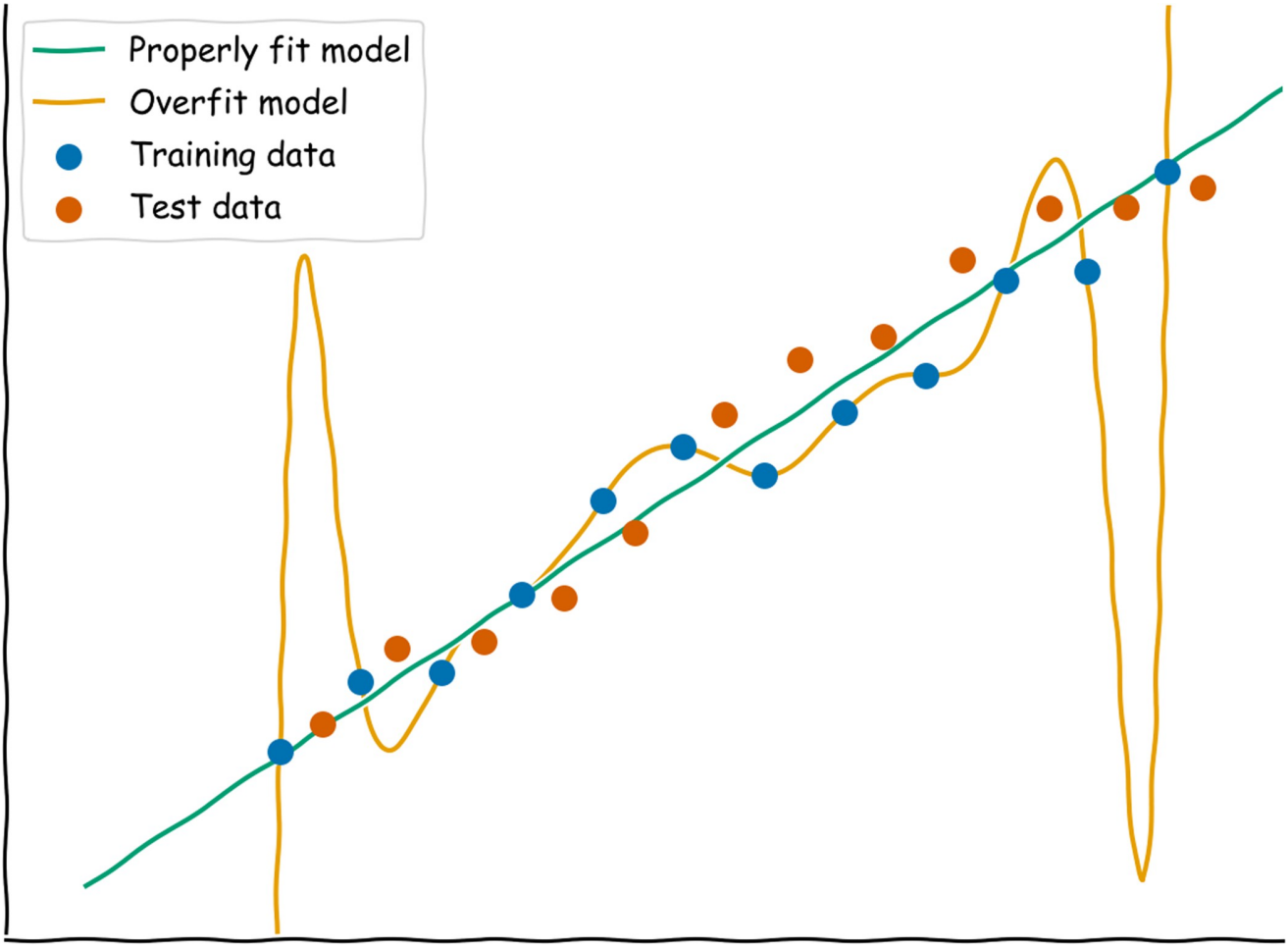


Fig 2. A visual example of overfitting and failure to generalize. While a high-degree polynomial achieves high accuracy on its training data, it performs poorly on the test data that have not been seen before. That is, the model has memorized the training dataset specifically rather than learning a generalizable pattern that represents data of this type. In contrast, a simple linear regression works equally well on both datasets.

https://doi.org/10.1371/journal.pcbi.1009803.g002

images were from and represents a type of artifact or batch effect that practitioners must be vigilant towards. When dealing with sequence data, holding out test data that are evolutionarily related or that share structural homology to the training data can result in overfitting that is hard to detect due to the inherent relatedness of the partitioned data [100]. In such situations, simply holding out test data selected from a random partition of the training data can be insufficient. The best remedy for identifying confounding variables is to know your data and to test models on truly independent data.

In essence, practitioners should split data into training, tuning, and single-use testing sets to assess the performance of the model on data that can provide a reliable estimate of its generalization performance. Furthermore, practitioners should be cognizant of the danger of skewed or biased data artificially inflating performance.

Tip 8: Deep learning models can be made more transparent

While model interpretability is a broad concept, in much of the machine learning literature, it refers to the ability to identify the discriminative features that influence or sway the

predictions. In certain cases, the goal behind interpretation is to understand the underlying data generating processes and biological mechanisms [101]. In other cases, the goal is to understand why a model made the prediction that it did for a specific example or set of examples. Machine learning models vary widely in terms of interpretability: Some are fully transparent while others are considered "black boxes" that make predictions with little ability to examine why. Logistic regression and decision tree models are generally considered interpretable [102]. In contrast, deep neural networks are often considered among the most difficult to interpret naively because they can have many parameters and nonlinear relationships.

Knowing which of the input variables influences a model's predictions, and potentially in what ways, can help with the application or extrapolation of machine learning models. This is particularly important in biomedicine. Subsequent decision-making often requires human input, and models are employed with the hope of better understanding why relationships exist in the first place. Furthermore, while prediction rules can be derived from high-throughput molecular datasets, most affordable clinical tests still rely on lower-dimensional measurements of a limited number of biomarkers. Therefore, it is often unclear how to translate the predictive capacity of deep learning models that encompass nonlinear relationships between countless input variables into clinically digestible terms. As a result, selecting which biomarkers to use for decision-making remains an important modeling and interpretation challenge. In fact, many authors attribute a lower uptake of deep learning tools in healthcare to interpretability challenges [103,104]. Nonetheless, strategies to interpret both machine learning and deep learning models are rapidly emerging, and the literature on the topic is growing quickly [105]. Instead of recommending specific methods for either deep learning–specific or general-purpose model interpretation, we suggest consulting a freely available and continually updated textbook [106].

Tip 9: Don't over interpret predictions

After training an accurate deep learning model, it is natural to want to use it to deduce relationships and inform scientific findings. However, be careful to interpret the model's predictions correctly. Given that deep learning models can be difficult to interpret intuitively, there is often a temptation to over interpret the predictions in indulgent or inaccurate ways. In accordance with the classic statistical saying "correlation doesn't imply causation," predictions by deep learning models rarely provide causal relationships. Accurately predicting an outcome does not mean that a causal mechanism has been learned, even when predictions are extremely accurate. In a poignant example, authors evaluated the capacities of several models to predict the probability of death for patients with pneumonia admitted to an intensive care unit [107,108]. The neural network model achieved the best predictive accuracy. However, after fitting a rule-based model to understand the relationships inherent to their data better, the authors discovered that the hospital data implied the rule "HasAsthma (x)⇒LowerRisk(x)." This rule contradicts medical understanding, as having asthma does not make pneumonia better! Nonetheless, the data supported this rule, as pneumonia patients with a history of asthma tended to receive more aggressive care. The neural network had, therefore, also learned to make predictions according to this rule despite the fact that it has nothing to do with causality or mechanism. Therefore, it would have been disastrous to guide treatment decisions according to the predictions of the neural network, even though the neural network had high predictive accuracy.

It is critical to avoid over interpreting deep learning models by viewing them for what they are: complex statistical models trained on high dimensional data. If causal inference is desired, special techniques for causal inference are required [109].

Tip 10: Actively consider the ethical implications of your work

While deep learning continues to be a powerful, transformative tool within life sciences research—spanning from basic biology and preclinical science to varied translational approaches and clinical studies—it is important to comment on ethical considerations. For instance, despite the fact that deep learning methods are helping to increase medical efficiency through improved diagnostic capability and risk assessment, certain biases may be inadvertently introduced into models related to patient age, race, and gender [110]. Deep learning practitioners may make use of datasets not representative of diverse populations and patient characteristics [111], thereby contributing to these problems.

Therefore, it is important to think thoroughly and cautiously about deep learning applications and their potential impact to persons and society. This includes being mindful of possible harms, injustices, and other types of wrongdoing. At a minimum, practitioners must ensure that, wherever relevant, their life sciences projects are fully compliant with local research governance and approval policies, legal requirements, Institutional Review Board policies, and any other relevant bodies and standards. Moreover, we offer below 3 tangible, action-oriented recommendations to further empower and enrichen deep learning researchers.

First, just as it is a best practice to keep a project-specific or programming-related issue tracker detailing known bugs and other technical issues, practitioners should get into the habit of keeping an active ethics register. In this register, ethical concerns can be raised, recorded, and resolved, exactly as software problems are triaged and fixed. Because projects using deep learning usually rely on writing code, an ethics register can be a part of the issue tracker in the version control system for the software itself. By colocating the two, practitioners can operationalize the concept that ethical problems are "bugs" that must be resolved, not nice-to-haves that can be considered at some indefinite point in the future. For practitioners intending to distribute trained models, having an ethics register can also facilitate creating a model card [112], a short document specifying the domains in which the model's performance was validated (for example, which model organism was used), how the performance was benchmarked, and known limitations and concerns. Second, to help foster a conscious ethics-oriented mindset, researchers should consider expanding journal clubs to include scholarly and popular articles detailing real-world ethics issues relevant to their scientific fields. This will help researchers to think more holistically and judiciously about their work and its implications. Third, we encourage individual- and team-level participation in professional societies [113] and other types of organizations [114] and events [115] related to the domains of artificial intelligence and data ethics as well as bioethics. This will encourage a sense of community and intellectual engagement and will keep practitioners abreast of cutting-edge insights and emerging professional standards.

Furthermore, practitioners may encounter datasets that cannot be shared, such as ones for which there would be significant ethical or legal issues associated with their release [116]. Examples of such data include classified or confidential data, biological data related to trade secrets, medical records, or other personally identifiable information [117]. While deep learning models can capture information-rich abstractions of multiple features of the data during the training process, these features may be more prone to leak the data that they were trained over if the model is shared or allowed to be queried with arbitrary inputs [118,119]. In other words, the complex relationships learned about the input data can potentially be used to infer characteristics about the original dataset, which reflects the facts that the strengths that imbue deep learning with its great predictive capacity may also raise the level of risk surrounding data privacy. Therefore, while there is tremendous promise for deep learning techniques to extract information that cannot readily be captured by traditional methods [120], it is imperative not

to share models trained on sensitive data. This also holds true for certain traditional machine learning methods that learn by capturing specific details of the full training data (for example, *k*-nearest neighbors models).

Techniques to train deep neural networks without sharing unencrypted access to data are being advanced through implementations of homomorphic encryption, which serves to enable equivalent prediction on data that is encrypted end-to-end [121,122]. Privacy-preserving techniques [123], such as differential privacy [124–126], can help to mitigate risks as long as the assumptions underlying these techniques are met. Other methods, such a distributed learning, in which small subsets of the data are processed independently in silos (possibly by different agents), are also promising but require careful investigation before applying to protected information [127]. While these methods provide a path toward a future where trained models and their predictions can be shared, more software development and theoretical advances will be required to make these techniques easy to apply correctly in many settings. Unless using these techniques, researchers must not share the weights or provide arbitrary access to the predictions of models trained on sensitive data.

Conclusions

Collectively, we have presented practical tips that emphasize cutting-edge insights and describe evolving professional standards. In addition, a number of our points are focused on safeguarding against the risks inherent to data science and deep learning. These risks include the over interpretation and misinterpretation of models, poor generalizability, and the potential to harm others. However, we want to strongly emphasize that when used ethically and responsibly, deep learning techniques have the potential to add tremendous value across a diverse range of contexts. After all, these techniques have already shown a remarkable capacity to meet or exceed the performance of both humans and traditional algorithms and have the potential to uncover biomedical insights that drive discovery and innovation. By taking a comprehensive and careful approach to deep learning based on critical thinking about research questions, planning to maintain rigor, and discerning how work might have far-reaching consequences with ethical dimensions, the life science community can advance reproducible, interpretable, and high-quality science that is enriching and beneficial for both scientists and society.

Acknowledgments

The authors would like the thank Daniel Himmelstein and the developers of Manubot for creating the software that enabled the collaborative composition of this manuscript. We would also like to thank Fábio Madeira (0000-0001-8728-9449), Victor Greiff (0000-0003-2622-5032), Shyam Saladi (0000-0001-9701-3059), Anshul Kundaje (0000-0003-3084-2287), Brett K. Beaulieu-Jones (0000-0002-6700-1468), Paul Brodersen (0000-0001-5216-7863), Michael M. Hoffman (0000-0002-4517-1562), and Isaac Lazzeri for their contributions to the discussions that comprised the initial stage of the drafting process.

References

1. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. Nat Rev Neurosci. 2020; 21:335–46. https://doi.org/10.1038/s41583-020-0277-3 PMID: 32303713
2. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018; 15:20170387. https://doi.org/10.1098/rsif.2017.0387 PMID: 29618526
3. Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for deep learning of molecular kinetics. Nat Commun. 2018; 9:5. https://doi.org/10.1038/s41467-017-02388-1 PMID: 29295994

4. Nielsen AAK, Voigt CA. Deep learning to predict the lab-of-origin of engineered DNA. Nat Commun. 2018; 9:3135. https://doi.org/10.1038/s41467-018-05378-z PMID: 30087331
5. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019; 25:60–4. https://doi.org/10.1038/s41591-018-0279-0 PMID: 30617323
6. Lee B. Benjamin-Lee/deep-rules GitHub repository. In: GitHub [Internet]. 2018. Available from: https://github.com/Benjamin-Lee/deep-rules
7. Himmelstein DS, Rubinetti V, Slochower DR, Hu D, Malladi VS, Greene CS, et al. Open collaborative writing with Manubot. PLoS Comput Biol. 2019; 15:e1007128. https://doi.org/10.1371/journal.pcbi.1007128 PMID: 31233491
8. Raschka S, Mirjalili V. Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. 3rd edition. Birmingham Mumbai: Packt; 2019.
9. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min. 2017; 10:35. https://doi.org/10.1186/s13040-017-0155-3 PMID: 29234465
10. Rudin C, Carlson D. The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis. arXiv. 2019 Jun. Report No.: 1906.01998. Available from: https://arxiv.org/abs/1906.01998
11. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2021. https://doi.org/10.1038/s41580-021-00407-0 PMID: 34518686
12. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. OMICS. 2018; 22:630–6. https://doi.org/10.1089/omi.2018.0097 PMID: 30124358
13. Mathew A, Amudha P, Sivakumari S. Deep Learning Techniques: An Overview. Advances in Intelligent Systems and Computing. Springer Science and Business Media LLC; 2021. https://doi.org/10.1007/978-981-15-3383-9_54
14. Raschka S, Patterson J, Nolet C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information. 2020; 11:193. https://doi.org/10.3390/info11040193
15. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Syst. 1989; 2:303–14. https://doi.org/10.1007/bf02551274
16. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Netw. 1991; 4:251–7. https://doi.org/10.1016/0893-6080(91)90009-t
17. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: Fast and Flexible Image Augmentations. Information. 2020; 11:125. https://doi.org/10.3390/info11020125
18. Ratner A, De Sa C, Wu S, Selsam D, Ré C. Data Programming: Creating Large Training Sets, Quickly. arXiv. 2016 May. Report No.: 1605.07723v3. Available from: https://arxiv.org/abs/1605.07723v3
19. Correa I, Drews P, Botelho S, de Souza MS, Tavano VM. Deep Learning for Microalgae Classification. Institute of Electrical and Electronics Engineers (IEEE). 2017. https://doi.org/10.1109/icmla.2017.0–183
20. Tian S, Lu S, Li C. WeText: Scene Text Detection under Weak Supervision. Institute of Electrical and Electronics Engineers (IEEE). 2017. https://doi.org/10.1109/iccv.2017.166
21. Flint C, Cearns M, Opel N, Redlich R, Mehler DMA, Emden D, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. Neuropsychopharmacology. 2021; 46:1510–7. https://doi.org/10.1038/s41386-021-01020-7 PMID: 33958703
22. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. Institute of Electrical and Electronics Engineers (IEEE). 2017. https://doi.org/10.1109/iccv.2017.97
23. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. Association for Computational Linguistics (ACL). 2020. https://doi.org/10.18653/v1/2020.acl-main.747
24. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv. 2016 Jan. Report No.: 1511.06348. Available from: https://arxiv.org/abs/1511.06348
25. Sze V, Chen Y-H, Yang T-J, Emer JS. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proc IEEE. 2017; 105:2295–329. https://doi.org/10.1109/jproc.2017.2761740
26. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv. 2020 Jul. Report No.: 2005.14165. Available from: https://arxiv.org/abs/2005.14165

27. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. arXiv. 2019 Jun. Report No.: 1906.02243. Available from: https://arxiv.org/abs/1906.02243

28. Madani A, McCann B, Naik N, Keskar NS, Anand N, Eguchi RR, et al. ProGen: Language Modeling for Protein Generation. arXiv. 2020 Apr. Report No.: 2004.03497. Available from: https://arxiv.org/abs/2004.03497

29. Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. arXiv. 2021 May. Report No.: 2007.06225. Available from: https://arxiv.org/abs/2007.06225

30. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci. 2021; 118: e2016239118. https://doi.org/10.1073/pnas.2016239118 PMID: 33876751

31. Alsouda Y, Pllana S, Kurti A. A Machine Learning Driven IoT Solution for Noise Classification in Smart Cities. arXiv. 2018 Sep. Report No.: 1809.00238. Available from: https://arxiv.org/abs/1809.00238

32. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensor Flow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv. 2016 Mar. Report No.: 1603.04467. Available from: https://arxiv.org/abs/1603.04467

33. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. 2019 Dec. Report No.: 1912.01703. Available from: https://arxiv.org/abs/1912.01703

34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12:2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html.

35. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. Lecture Notes in Computer Science. Springer Science and Business Media LLC; 2016. https://doi.org/10.1007/978-3-319-31204-0_9

36. GitHub—apple/turicreate: Turi Create simplifies the development of custom machine learning models. In: GitHub [Internet]. [cited 2021 Nov 29]. Available from: https://github.com/apple/turicreate

37. Jin H, Song Q, Hu X. Auto-Keras: An Efficient Neural Architecture Search System. arXiv. 2019 Mar. Report No.: 1806.10282. Available from: https://arxiv.org/abs/1806.10282

38. Keras: the Python deep learning API. [cited 2021 Nov 29]. Available from: https://keras.io/

39. Howard J, Gugger S. Fastai: A Layered API for Deep Learning. Information. 2020; 11:108. https://doi.org/10.3390/info11020108

40. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017; 60:84–90. https://doi.org/10.1145/3065386

41. Lin E, Kuo P-H, Liu Y-L, Yu YW-Y, Yang AC, Tsai S-J. A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers. Front Psychiatry. 2018; 9:290. https://doi.org/10.3389/fpsyt.2018.00290 PMID: 30034349

42. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol. 2018; 36:257–72. https://doi.org/10.1007/s11604-018-0726-3 PMID: 29498017

43. Rivenson Y, Göröcs Z, Günaydin H, Zhang Y, Wang H, Ozcan A. Deep learning microscopy. Optica. 2017; 4:1437. https://doi.org/10.1364/optica.4.001437

44. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc. 2017; 24:813–21. https://doi.org/10.1093/jamia/ocw180 PMID: 28339747

45. Ferreira AC, Silva LR, Renna F, Brandl HB, Renoult JP, Farine DR, et al. Deep learning-based methods for individual recognition in small birds. Methods Ecol Evol. 2020; 11:1072–85. https://doi.org/10.1111/2041-210x.13436

46. Oussidi A, Elhassouny A. Deep generative models: Survey. Institute of Electrical and Electronics Engineers (IEEE). 2018. https://doi.org/10.1109/isacv.2018.8354080

47. Lee AJ, Park Y, Doing G, Hogan DA, Greene CS. Correcting for experiment-specific variability in expression compendia can remove underlying signals. GigaScience. 2020; 9:giaa117. https://doi.org/10.1093/gigascience/giaa117 PMID: 33140829

48. Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D. Deep Reinforcement Learning that Matters. arXiv. 2019 Jan. Report No.: 1709.06560. Available from: https://arxiv.org/abs/1709.06560

49. Zhou Z, Kearnes S, Li L, Zare RN, Riley P. Optimization of Molecules via Deep Reinforcement Learning. Sci Rep. 2019; 9:10752. https://doi.org/10.1038/s41598-019-47148-x PMID: 31341196

50. Fu W, Menzies T. Easy over hard: a case study on deep learning. Association for Computing Machinery (ACM). 2017. https://doi.org/10.1145/3106237.3106256

51. Smith AM, Walsh JR, Long J, Davis CB, Henstock P, Hodge MR, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC Bioinformatics. 2020; 21:119. https://doi.org/10.1186/s12859-020-3427-8 PMID: 32197580

52. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018; 1:18. https://doi.org/10.1038/s41746-018-0029-1 PMID: 31304302

53. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyperparameters and comparison of performance to shallow methods for modeling bioactivity data. J Chem. 2017; 9:42. https://doi.org/10.1186/s13321-017-0226-y PMID: 29086090

54. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med. 2019; 2:43. https://doi.org/10.1038/s41746-019-0122-0 PMID: 31304389

55. Papernot N, McDaniel P. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. arXiv. 2018 Mar. Report No.: 1803.04765. Available from: https://arxiv.org/abs/1803.04765

56. Jiang H, Kim B, Guan MY, Gupta M. To Trust Or Not To Trust A Classifier. arXiv. 2018 Oct. Report No.: 1805.11783. Available from: https://arxiv.org/abs/1805.11783

57. Hu Q, Greene CS. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. Pac Symp Biocomput. 2018. https://doi.org/10.1142/9789813279827_0033

58. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost F Da V, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput Biol. 2016; 12:e1004947. https://doi.org/10.1371/journal.pcbi.1004947 PMID: 27415786

59. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol. 2017; 35:342–6. https://doi.org/10.1038/nbt.3780 PMID: 28288103

60. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol. 2013; 9:e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232

61. Rule A, Birmingham A, Zuniga C, Altintas I, Huang S-C, Knight R, et al. Ten Simple Rules for Reproducible Research in Jupyter Notebooks. arXiv. 2018 Oct. Report No.: 1810.08055. Available from: https://arxiv.org/abs/1810.08055

62. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021; 18:1132–5. https://doi.org/10.1038/s41592-021-01256-7 PMID: 34462593

63. NVIDIA. Deep Learning SDK Documentation. 1 Nov 2018. Available from: https://docs.nvidia.com/deeplearning/sdk/cudnn-developer-guide/index.html#reproducibility

64. Wilson AC, Roelofs R, Stern M, Srebro N, Recht B. The Marginal Value of Adaptive Gradient Methods in Machine Learning. Adv Neural Inf Proces Syst. 2017; 30. Available from: https://papers.nips.cc/paper/2017/hash/81b3833e2504647f9d794f7d7b9bf341-Abstract.html

65. Vision T. The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem. Nat Prec. 2010. https://doi.org/10.1038/npre.2010.4595.1

66. Singh J. Fig Share. J Pharmacol Pharmacother. 2011; 2:138. https://doi.org/10.4103/0976-500X.81919 PMID: 21772785

67. Dillen M, Groom Q, Agosti D, Nielsen L. Zenodo, an Archive and Publishing Repository: A tale of two herbarium specimen pilot projects. Biodivers Inf Sci Stand. 2019; 3:e37080. https://doi.org/10.3897/biss.3.37080

68. Foster MSLSED, Deardorff MLISA. Open Science Framework (OSF). J Med Libr Assoc. 2017;105. https://doi.org/10.5195/jmla.2017.88

69. Gundersen OE, Gil Y, Aha DW. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. AI Mag. 2018; 39:56–68. https://doi.org/10.1609/aimag.v39i3.2816

70. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001; 29:365–71. https://doi.org/10.1038/ng1201-365 PMID: 11726920

71. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–9. https://doi.org/10.1038/nrg2825 PMID: 20838408

72. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. Springer Science and Business Media LLC; 2012. https://doi.org/10.1007/978-3-642-35289-8

73. Bai S, Kolter JZ, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv. 2018 Apr. Report No.: 1803.01271. Available from: https://arxiv.org/abs/1803.01271

74. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv. 2017 Dec. Report No.: 1706.03762. Available from: https://arxiv.org/abs/1706.03762

75. Raschka S, Kaufman B. Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition. Methods. 2020; 180:89–110. https://doi.org/10.1016/j.ymeth.2020.06.016 PMID: 32645448

76. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436–44. https://doi.org/10.1038/nature14539 PMID: 26017442

77. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2. Cambridge, MA, USA: MIT Press; 2014. p. 3320–3328. Available from: https://dl.acm.org/doi/abs/10.5555/2969033.2969197

78. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015; 115:211–52. https://doi.org/10.1007/s11263-015-0816-y

79. Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. J Digit Imaging. 2016; 30:95–101. https://doi.org/10.1007/s10278-016-9914-9 PMID: 27730417

80. Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat Biotechnol. 2019; 37:592–600. https://doi.org/10.1038/s41587-019-0140-0 PMID: 31138913

81. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. Association for Computational Linguistics (ACL). 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6

82. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2022; 3:1–23. https://doi.org/10.1145/3458754

83. Chithrananda S, Grand G, Ramsundar B. Chem BERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv. 2020 Oct. Report No.: 2010.09885. Available from: https://arxiv.org/abs/2010.09885

84. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. Institute of Electrical and Electronics Engineers (IEEE). 2014. https://doi.org/10.1109/cvprw.2014.131

85. Zheng X, Wang Y, Wang G, Liu J. Fast and robust segmentation of white blood cell images by self-supervised learning. Micron. 2018; 107:55–71. https://doi.org/10.1016/j.micron.2018.01.010 PMID: 29425969

86. Zhang W, Li R, Zeng T, Sun Q, Kumar S, Ye J, et al. Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. IEEE Transactions on Big Data. 2020; 6:322–33. https://doi.org/10.1109/tbdata.2016.2573280

87. Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. 1993; 6:861–7. https://doi.org/10.1016/s0893-6080(05)80131-5

88. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014; 15:1929–58. Available from: http://dl.acm.org/citation.cfm?id=2670313.

89. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37. Lille, France: JMLR.org; 2015. p. 448–456. Available from: https://dl.acm.org/citation.cfm?id=3045118.3045167

90. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci. 2019; 116:15849–54. https://doi.org/10.1073/pnas.1903070116 PMID: 31341078

91. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv. 2020 Nov. Report No.: 1811.12808. Available from: https://arxiv.org/abs/1811.12808

92. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998; 10:1895–923. https://doi.org/10.1162/089976698300017197 PMID: 9744903

93. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014; 15:1929–58. Available from: http://jmlr.org/papers/v15/srivastava14a.html.

94. Krogh A, Hertz JA. A simple weight decay can improve generalization. Proceedings of the 4th International Conference on Neural Information Processing Systems. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1991. p. 950–957. Available from: http://dl.acm.org/citation.cfm?id=2986916.2987033

95. Chuang KV, Keiser MJ. Adversarial Controls for Scientific Machine Learning. ACS Chem Biol. 2018; 13:2819–21. https://doi.org/10.1021/acschembio.8b00881 PMID: 30336670

96. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015; 10:e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806

97. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol Pharm. 2017; 14:4462–75. https://doi.org/10.1021/acs.molpharmaceut.7b00578 PMID: 29096442

98. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Association for Computing Machinery (ACM). 2006. https://doi.org/10.1016/j.ijmedinf.2005.07.035 PMID: 16125999

99. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018; 15:e1002683. https://doi.org/10.1371/journal.pmed.1002683 PMID: 30399157

100. Walsh I, Pollastri G, Tosatto SCE. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016; 17:831–40. https://doi.org/10.1093/bib/bbv082 PMID: 26411473

101. Bemister-Buffington J, Wolf AJ, Raschka S, Kuhn LA. Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition. Biomol Ther. 2020; 10:454. https://doi.org/10.3390/biom10030454 PMID: 32183371

102. Raschka S, Scott AM, Huertas M, Li W, Kuhn LA. Automated Inference of Chemical Discriminants of Biological Activity. Methods in Molecular Biology. Springer Science and Business Media LLC; 2018. https://doi.org/10.1007/978-1-4939-7756-7_16 PMID: 29594779

103. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017; 21:4–21. https://doi.org/10.1109/JBHI.2016.2636665 PMID: 28055930

104. Towards trustable machine learning. Nat Biomed Eng. 2018; 2:709–10. https://doi.org/10.1038/s41551-018-0315-x PMID: 31015650

105. Fan F, Xiong J, Li M, Wang G. On Interpretability of Artificial Neural Networks: A Survey. arXiv. 2021 Sep. Report No.: 2001.02522. Available from: https://arxiv.org/abs/2001.02522

106. Molnar C. Interpretable Machine Learning. Available from: https://christophm.github.io/interpretable-ml-book/

107. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med. 1997; 9:107–38. https://doi.org/10.1016/s0933-3657(96)00367-3 PMID: 9040894

108. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare. Association for Computing Machinery (ACM). 2015. https://doi.org/10.1145/2783258.2788613

109. Luo Y, Peng J, Ma J. When causal inference meets deep learning. Nat Mach Intell. 2020; 2:426–7. https://doi.org/10.1038/s42256-020-0218-x

110. Ho A. Deep Ethical Learning: Taking the Interplay of Human and Artificial Intelligence Seriously. Hastings Cent Rep. 2019; 49:36–9. https://doi.org/10.1002/hast.977 PMID: 30790317

111. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. Health Aff. 2014; 33:1139–47. https://doi.org/10.1377/hlthaff.2014.0048 PMID: 25006139

112. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. Association for Computing Machinery (ACM). 2019. https://doi.org/10.1145/3287560.3287596

113. American Society for Bioethics and Humanities. [cited 2021 Nov 29]. Available from: https://asbh.org/

114. 10 organizations leading the way in ethical AI—SAGE Ocean | Big Data, New Tech, Social Science. 2021 Jan 12 [cited 2021 Nov 29]. Available from: https://web.archive.org/web/20210112231619/https://ocean.sagepub.com/blog/10-organizations-leading-the-way-in-ethical-ai

115. Artificial Intelligence, Ethics, and Society—Home. [cited 2021 Nov 29]. Available from: https://www.aies-conference.com/2021/

116. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. PLoS Comput Biol. 2017; 13:e1005399. https://doi.org/10.1371/journal.pcbi.1005399 PMID: 28358831

117. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020; 21:615–29. https://doi.org/10.1038/s41576-020-0257-5 PMID: 32694666

118. Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Association for Computing Machinery (ACM). 2015. https://doi.org/10.1145/2810103.2813677

119. Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks against Machine Learning Models. arXiv. 2017 Apr. Report No.: 1610.05820. Available from: https://arxiv.org/abs/1610.05820

120. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv. 2015 Nov. Report No.: 1509.09292. Available from: https://arxiv.org/abs/1509.09292

121. Titus AJ, Flower A, Hagerty P, Gamble P, Lewis C, Stavish T, et al. SIG-DB: Leveraging homomorphic encryption to securely interrogate privately held genomic databases. PLoS Comput Biol. 2018; 14: e1006454. https://doi.org/10.1371/journal.pcbi.1006454 PMID: 30180163

122. Badawi AA, Chao J, Lin J, Mun CF, Sim JJ, Tan BHM, et al. Towards the AlexNet Moment for Homomorphic Encryption: HCNN, theFirst Homomorphic CNN on Encrypted Data with GPUs. arXiv. 2020 Aug. Report No.: 1811.00778. Available from: https://arxiv.org/abs/1811.00778

123. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, et al. A generic framework for privacy preserving deep learning. arXiv. 2018 Nov. Report No.: 1811.04017. Available from: https://arxiv.org/abs/1811.04017

124. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep Learning with Differential Privacy. Association for Computing Machinery (ACM). 2016. https://doi.org/10.1145/2976749.2978318

125. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. Circ Cardiovasc Qual Outcomes. 2019; 12. https://doi.org/10.1161/CIRCOUTCOMES.118.005122 PMID: 31284738

126. Beaulieu-Jones BK, Yuan W, Finlayson SG, Wu ZS. Privacy-Preserving Distributed Deep Learning for Clinical Data. arXiv. 2018 Dec. Report No.: 1812.01484. Available from: https://arxiv.org/abs/1812.01484

127. Zerka F, Urovi V, Bottari F, Leijenaar RTH, Walsh S, Gabrani-Juma H, et al. Privacy preserving distributed learning classifiers–Sequential learning with small sets of data. Comput Biol Med. 2021; 136:104716. https://doi.org/10.1016/j.compbiomed.2021.104716 PMID: 34364262

D.13 Image analysis and models of cells

Further explanations for Chapter 13.

D.13.1 Images as data

- Bankhead (2022) — An open textbook on what digital images are, how they are processed and measured, and how to avoid common mistakes; the best starting point for the classical part of image analysis. *(book chapter; CC BY 4.0)*

  Introduction to Bioimage Analysis. Bankhead (2022). *Online book*. https://bioimagebook.github.io/

- Moore et al. (2023) — Explains why large image data need new file formats and metadata standards, and how OME-Zarr is used to share them. *(review; CC BY 4.0)*

  OME-Zarr: a cloud-optimized bioimaging file format with international community support. Moore et al. (2023). *Histochemistry and Cell Biology*. doi:10.1007/s00418-023-02209-1

  > A growing community is constructing a next-generation file format (NGFF) for bioimaging to overcome problems of scalability and heterogeneity. Organized by the Open Microscopy Environment (OME), individuals and institutes across diverse modalities facing these problems have designed a format specification process (OME-NGFF) to address these needs. This paper brings together a wide range of those community members to describe the cloud-optimized format itself-OME-Zarr-along with tools and data resources available today to increase FAIR access and remove barriers in the scientific process. The current momentum offers an opportunity to unify a key component of the bioimaging domain-the file format that underlies so many personal, institutional, and global data management and analysis tasks.

D.13.2 Deep learning for images

- Hallou et al. (2021) — A clear introduction to classification, segmentation, tracking and restoration with deep learning, written for biologists, with practical advice on training data and validation. *(review; CC BY 4.0)*

  Deep learning for bioimage analysis in developmental biology. Hallou et al. (2021). *Development*. doi:10.1242/dev.199616

  > Deep learning has transformed the way large and complex image datasets can be processed, reshaping what is possible in bioimage analysis. As the complexity and size of bioimage data continues to grow, this new analysis paradigm is becoming increasingly ubiquitous. In this Review, we begin by introducing the concepts needed for beginners to understand deep learning. We then review how deep learning has impacted bioimage analysis and explore the open-source resources available to integrate it into a research project. Finally, we discuss the future of deep learning applied to cell and developmental biology. We analyze how state-of-the-art methodologies have the potential to transform our understanding of biological systems through new image-based analysis and modelling that integrate multimodal inputs in space and time.

- Chamier et al. (2021) — Describes notebooks for training common networks (U-Net, StarDist, CARE and others) without installing software; useful for trying the methods yourself. *(tutorial; CC BY 4.0)*

  Democratising deep learning for microscopy with ZeroCostDL4Mic. von Chamier et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-22518-0

> Deep Learning (DL) methods are powerful analytical tools for microscopy and can outperform conventional image processing pipelines. Despite the enthusiasm and innovations fuelled by DL technology, the need to access powerful and compatible resources to train DL networks leads to an accessibility barrier that novice users often find difficult to overcome. Here, we present ZeroCostDL4Mic, an entry-level platform simplifying DL access by leveraging the free, cloud-based computational resources of Google Colab. ZeroCostDL4Mic allows researchers with no coding expertise to train and apply key DL networks to perform tasks including segmentation (using U-Net and StarDist), object detection (using YOLOv2), denoising (using CARE and Noise2Void), super-resolution microscopy (using Deep-STORM), and image-to-image translation (using Label-free prediction - fnet, pix2pix and CycleGAN). Importantly, we provide suitable quantitative tools for each network to evaluate model performance, allowing model optimisation. We demonstrate the application of the platform to study multiple biological processes.

D.13.3 Image-based profiling

- Caicedo et al. (2017) — The standard description of the profiling workflow, from illumination correction and feature extraction to normalisation, batch correction and similarity measures. *(review; CC BY 4.0)*

  Data-analysis strategies for image-based cell profiling. Caicedo et al. (2017). *Nature Methods*. doi:10.1038/nmeth.4397

  > Image-based cell profiling is a high-throughput strategy for the quantification of phenotypic differences among a variety of cell populations. It paves the way to studying biological systems on a large scale by using chemical and genetic perturbations. The general workflow for this technology involves image acquisition with high-throughput microscopy systems and subsequent image processing and analysis. Here, we introduce the steps required to create high-quality image-based (i.e., morphological) profiles from a collection of microscopy images. We recommend techniques that have proven useful in each stage of the data analysis process, on the basis of the experience of 20 laboratories worldwide that are refining their image-based cell-profiling methodologies in pursuit of biological discovery. The recommended techniques cover alternatives that may suit various biological goals, experimental designs, and laboratories' preferences.

- Stirling et al. (2021) — Describes the most widely used software for extracting features from Cell Painting and other screens. *(tutorial; CC BY 4.0)*

  CellProfiler 4: improvements in speed, utility and usability. Stirling et al. (2021). *BMC Bioinformatics*. doi:10.1186/s12859-021-04344-9

  > Background Imaging data contains a substantial amount of information which can be difficult to evaluate by eye. With the expansion of high throughput microscopy methodologies producing increasingly large datasets, automated and objective analysis of the resulting images is essential to effectively extract biological information from this data. CellProfiler is a free, open source image analysis program which enables researchers to generate modular pipelines with which to process microscopy images into interpretable measurements. Results Herein we describe CellProfiler 4, a new version of this software with expanded functionality. Based on user feedback, we have made several user interface refinements to improve the usability of the software. We introduced new modules to expand the capabilities of the software. We also evaluated performance and made targeted optimizations to reduce the time and cost associated with running common large-scale analysis pipelines. Conclusions CellProfiler 4 provides significantly improved performance in complex workflows compared to previous versions. This release will ensure that researchers will have continued access to CellProfiler's powerful computational tools in the coming years.

D.13.4 Protein localisation from images

- Xiao et al. (2024) — Reviews both sequence-based and image-based prediction of localisation, and so connects this section with the sequence-based methods of chapter 10. *(review; CC BY 4.0)*

  A Review for Artificial Intelligence Based Protein Subcellular Localization. Xiao et al. (2024). *Biomolecules*. doi:10.3390/biom14040409

> Proteins need to be located in appropriate spatiotemporal contexts to carry out their diverse biological functions. Mislocalized proteins may lead to a broad range of diseases, such as cancer and Alzheimer's disease. Knowing where a target protein resides within a cell will give insights into tailored drug design for a disease. As the gold validation standard, the conventional wet lab uses fluorescent microscopy imaging, immunoelectron microscopy, and fluorescent biomarker tags for protein subcellular location identification. However, the booming era of proteomics and high-throughput sequencing generates tons of newly discovered proteins, making protein subcellular localization by wet-lab experiments a mission impossible. To tackle this concern, in the past decades, artificial intelligence (AI) and machine learning (ML), especially deep learning methods, have made significant progress in this research area. In this article, we review the latest advances in AI-based method development in three typical types of approaches, including sequence-based, knowledge-based, and image-based methods. We also elaborately discuss existing challenges and future directions in AI-based method development in this research field.

- Kobayashi et al. (2022) — A readable example of how self-supervised learning on images groups proteins by localisation without labels (cytoself). *(primary study; CC BY 4.0)*

  Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Kobayashi et al. (2022). *Nature Methods*. doi:10.1038/s41592-022-01541-z

  > Explaining the diversity and complexity of protein localization is essential to fully understand cellular architecture. Here we present cytoself, a deep-learning approach for fully self-supervised protein localization profiling and clustering. Cytoself leverages a self-supervised training scheme that does not require preexisting knowledge, categories or annotations. Training cytoself on images of 1,311 endogenously labeled proteins from the OpenCell database reveals a highly resolved protein localization atlas that recapitulates major scales of cellular organization, from coarse classes, such as nuclear and cytoplasmic, to the subtle localization signatures of individual protein complexes. We quantitatively validate cytoself's ability to cluster proteins into organelles and protein complexes, showing that cytoself outperforms previous self-supervised approaches. Moreover, to better understand the inner workings of our model, we dissect the emergent features from which our clustering is derived, interpret them in the context of the fluorescence images, and analyze the performance contributions of each component of our approach.

D.13.5 Tissues: multiplexed imaging and computational pathology

- Park et al. (2022) — An overview of imaging- and sequencing-based spatial methods for RNA, proteins and other molecules, and how they are combined. *(review; CC BY 4.0)*

  Spatial omics technologies at multimodal and single cell/subcellular level. Park et al. (2022). *Genome Biology*. doi:10.1186/s13059-022-02824-6

  > Spatial omics technologies enable a deeper understanding of cellular organizations and interactions within a tissue of interest. These assays can identify specific compartments or regions in a tissue with differential transcript or protein abundance, delineate their interactions, and complement other methods in defining cellular phenotypes. A variety of spatial methodologies are being developed and commercialized; however, these techniques differ in spatial resolution, multiplexing capability, scale/throughput, and coverage. Here, we review the current and prospective landscape of single cell to subcellular resolution spatial omics technologies and analysis tools to provide a comprehensive picture for both research and clinical applications.

- Howard et al. (2021) — Shows how hidden differences between hospitals inflate the apparent accuracy of pathology models; a good case study of data leakage. *(primary study; CC BY 4.0)*

  The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Howard et al. (2021). *Nature Communications*. doi:10.1038/s41467-021-24698-1

  > The Cancer Genome Atlas (TCGA) is one of the largest biorepositories of digital histology. Deep learning (DL) models have been trained on TCGA to predict numerous features directly from histology, including survival, gene expression patterns, and driver mutations. However, we demonstrate that these features vary substantially across tissue submitting sites in TCGA for over 3,000 patients with six cancer subtypes. Additionally, we show that histologic image differences between submitting sites can easily be identified with DL. Site detection remains possible despite commonly used color normalization and augmentation

methods, and we quantify the image characteristics constituting this site-specific digital histology signature. We demonstrate that these site-specific signatures lead to biased accuracy for prediction of features including survival, genomic mutations, and tumor stage. Furthermore, ethnicity can also be inferred from site-specific signatures, which must be accounted for to ensure equitable application of DL. These site-specific signatures can lead to overoptimistic estimates of model performance, and we propose a quadratic programming method that abrogates this bias by ensuring models are not trained and validated on samples from the same site.

D.13.6 Cell atlases and virtual cells

- Regev et al. (2017) — The white paper that defined the aims of the Human Cell Atlas, including what is meant by a cell type and why images and locations are part of it. *(perspective; CC BY 4.0)*

  The Human Cell Atlas. Regev et al. (2017). *eLife*. doi:10.7554/elife.27041

  > The recent advent of methods for high-throughput single-cell molecular profiling has catalyzed a growing sense in the scientific community that the time is ripe to complete the 150-year-old effort to identify all cell types in the human body. The Human Cell Atlas Project is an international collaborative effort that aims to define all human cell types in terms of distinctive molecular profiles (such as gene expression profiles) and to connect this information with classical cellular descriptions (such as location and morphology). An open comprehensive reference map of the molecular state of cells in healthy human tissues would propel the systematic study of physiological states, developmental trajectories, regulatory circuitry and interactions of cells, and also provide a framework for understanding cellular dysregulation in human disease. Here we describe the idea, its potential utility, early proofs-of-concept, and some design considerations for the Human Cell Atlas, including a commitment to open data, code, and community.

- Kedzierska et al. (2025) — A careful evaluation of single-cell foundation models against simple baselines; useful for discussing what virtual cell models can and cannot yet do. *(primary study; CC BY 4.0)*

  Zero-shot evaluation reveals limitations of single-cell foundation models. Kedzierska et al. (2025). *Genome Biology*. doi:10.1186/s13059-025-03574-x

  > Foundation models such as scGPT and Geneformer have not been rigorously evaluated in a setting where they are used without any further training (i.e., zero-shot). Understanding the performance of models in zero-shot settings is critical to applications that exclude the ability to fine-tune, such as discovery settings where labels are unknown. Our evaluation of the zero-shot performance of Geneformer and scGPT suggests that, in some cases, these models may face reliability challenges and could be outperformed by simpler methods. Our findings underscore the importance of zero-shot evaluations in development and deployment of foundation models in single-cell research.

D.13.7 Whole-cell models and cryo-electron tomography

- Young and Villa (2023) — An open review of cryo-electron tomography, from sample thinning by focused ion beams to subtomogram averaging and visual proteomics. *(review; CC BY 4.0)*

  Bringing Structure to Cell Biology with Cryo-Electron Tomography. Young & Villa (2023). *Annual Review of Biophysics*. doi:10.1146/annurev-biophys-111622-091327

  > Recent advances in cryo-electron microscopy have marked only the beginning of the potential of this technique. To bring structure into cell biology, the modality of cryo-electron tomography has fast developed into a bona fide in situ structural biology technique where structures are determined in their native environment, the cell. Nearly every step of the cryo-focused ion beam-assisted electron tomography (cryo-FIB-ET) workflow has been improved upon in the past decade, since the first windows were carved into cells, unveiling macromolecular networks in near-native conditions. By bridging structural and cell biology, cryo-FIB-ET is advancing our understanding of structure-function relationships in their native environment and becoming a tool for discovering new biology.

- Stevens et al. (2023) — Explains how a whole minimal cell could be built and simulated with coarse-grained molecular dynamics, and what data this requires. *(perspective; CC BY 4.0)*

  Molecular dynamics simulation of an entire cell. Stevens et al. (2023). *Frontiers in Chemistry*. doi:10.3389/fchem.2023.1106495

  > The ultimate microscope, directed at a cell, would reveal the dynamics of all the cell's components with atomic resolution. In contrast to their real-world counterparts, computational microscopes are currently on the brink of meeting this challenge. In this perspective, we show how an integrative approach can be employed to model an entire cell, the minimal cell, JCVI-syn3A, at full complexity. This step opens the way to interrogate the cell's spatio-temporal evolution with molecular dynamics simulations, an approach that can be extended to other cell types in the near future.

The following open-access paper(s) are reproduced under their original licences.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Annu Rev Biophys. Author manuscript; available in PMC 2024 January 03.

Published in final edited form as:

Annu Rev Biophys. 2023 May 09; 52: 573–595. doi:10.1146/annurev-biophys-111622-091327.

Bringing Structure to Cell Biology with Cryo-Electron Tomography

Lindsey N. Young[1], **Elizabeth Villa**[1,2]

[1]Department of Molecular Biology, University of California, San Diego, La Jolla, California, USA;

[2]Howard Hughes Medical Institute, University of California, San Diego, La Jolla, California, USA

Abstract

Recent advances in cryo-electron microscopy have marked only the beginning of the potential of this technique. To bring structure into cell biology, the modality of cryo-electron tomography has fast developed into a bona fide in situ structural biology technique where structures are determined in their native environment, the cell. Nearly every step of the cryo-focused ion beam-assisted electron tomography (cryo-FIB-ET) workflow has been improved upon in the past decade, since the first windows were carved into cells, unveiling macromolecular networks in near-native conditions. By bridging structural and cell biology, cryo-FIB-ET is advancing our understanding of structure–function relationships in their native environment and becoming a tool for discovering new biology.

Keywords

cryo-electron tomography; in situ structural biology; quantitative cell biology; biophysical modeling; molecular architecture

INTRODUCTION

Since the 1940s, direct visualization into the inner working of a cell by conventional transmission electron microscopy (TEM) has facilitated discoveries and generated both new hypotheses and awe as researchers peered directly into the complex and heterogeneous cellular environment for the first time (154). TEM experiments on cells that were fixed, permeabilized, and stained, a process referred to as conventional TEM, gave way to high-pressure freezing (HPF) methods and freeze-substitution, which allowed researchers to inch closer to native conditions and better visualize a cell's interior. Separately, the development of direct electron detector devices (DDDs) and new image processing algorithms precipitated the resolution revolution in single-particle cryo-electron microscopy (cryo-EM) (81). These technical developments further facilitated high-resolution 3D reconstructions of in vitro samples determined, not by combining thousands of 2D

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See credit lines of images or other third-party material in this article for license information.

evilla@ucsd.edu .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

projections of different particles of the same kind (single-particle cryo-EM), but from projections acquired at multiple orientations or tilts of the same sample (tomography) (101, 127). The next hurdle was to visualize these macromolecules directly in cells (in situ or in cellulo), without any fixation, sectioning, or other perturbations. How could one achieve this? By borrowing techniques from materials science and by operating under cryogenic temperatures, researchers carved out a whole new field in which they could open windows directly into the complex cellular environment using focused ion beam (FIB) milling (94). The preservation of high-resolution information created a new field of structural cell biology studies in which subtomogram analysis, combined with other methods such as integrative modeling, has expanded how structure–function relationships can be explored and investigated (8, 94).

BACKGROUND

Investigating how individual macromolecules organize to perform all functions in the cell, e.g., cellular growth, division, and migration, requires understanding how individual small players act in concert, resulting in emergent properties that accomplish larger tasks. Macromolecular complexes are roughly tens of nanometers across (102), and their interactions are governed by chemical bonds, electrostatic interactions, hydrogen bonds, and van der Waals forces. These proteins and macromolecular complexes then assemble to form structures that are micrometers in size (e.g., organelles, filaments). Investigating structure–function relationships in situ requires the ability to image across multiple length scales. Investigating how these distinct structures function in their native environment requires the ability to visualize macromolecules that are minimally perturbed. By imaging across multiple length scales (from angstroms to microns), an electron microscopist can investigate assemblies at the molecular level and also define relevant spatial relationships—all within a single cell. Furthermore, new discoveries can be made via electron microscopy (EM) because the interaction between the electron beam and the sample occurs for all atoms in the sample that are subsequently captured in the resulting image. This is in contrast to fluorescence microscopy, in which prior knowledge and labeling of key macromolecules are required for further investigation of their molecular function, and only labeled molecules are captured in the image. Essentially, an electron beam sees all.

Cryo-electron tomography (cryo-ET) is a modality of TEM in which a flash-frozen sample is imaged at multiple orientations by tilting the sample. The resulting collection of images, or tilt series, is aligned to reconstruct the sample in three dimensions into a tomogram, where the intensities in a sample are roughly proportional to the mass of the underlying atoms (159). From the tomogram, one can extract specific regions containing macromolecules of interest. These subtomograms can then be extracted, aligned, averaged, and further analyzed to obtain 3D structures of molecular complexes, a technique called subtomogram analysis (STA). In a proof of the principle that high-resolution information can be recovered from in situ data, a recent cryo-ET reconstruction of a ribosome stalled during protein translation was resolved to 3.4 Å (137). However, as only transmitted electrons are detected, there is a strict thickness limitation to this methodology. The mean free path of 300 keV electrons was determined to be 280 nm for free vesicles in cryogenic conditions (66, 97, 119, 162). Larger thicknesses result in multiple interactions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of the electrons with the sample, dramatically reducing the signal-to-noise ratio (SNR) in the images. As most cells are thicker than 500 nm, samples need to be thinned. In the past decade, this limitation was overcome by cryo-FIB milling (141, 144). This technique ablates away most of the cellular material, leaving behind a window of several to tens of microns in *x* and *y*, with a thickness of approximately 80–300 nm. Cryo-FIB-ET is the highest-resolution technique available to image macromolecules inside cells, and an ever increasing number of landmark discoveries have been made in the past decade using this technology (1, 3, 47, 48). In this review, we describe the cryo-FIB-ET workflow, the recent advances in the field (sample preparation, data acquisition, and analysis),the integration with other methods, considerations for designing a successful project, and an outlook for the field.

THE CRYO-FOCUSED ION BEAM MILLING AND ELECTRON TOMOGRAPHY WORKFLOW

Most commonly, cultured cells are grown or deposited onto an EM grid, and the sample is then rapidly frozen into a noncrystalline, glass-like vitrified state by plunging it into liquid ethane cooled to liquid nitrogen temperature. The result is preservation of macromolecular structures in their cellular networks in near-native conditions (40, 72). The sample is then transferred under cryogenic conditions (liquid nitrogen temperature) to a dual beam microscope equipped with an FIB and a scanning electron microscope (SEM); the FIB-SEM is routinely used within materials science for microstructural analysis, chip editing, nanomachining, and sample preparation for TEM (61). For vitrified cells, regions above and below the region of interest are selectively ablated as the FIB rasters across, removing cellular material via the interaction between an ion source, typically of gallium, and the sample in a process referred to as milling (94, 145). The resulting lamella is milled to a thickness of 80–300 nm (94, 145). The sample is transferred to a TEM, and regions of interest within the lamellae are imaged at different tilt angles to collect a tilt series that is later aligned and reconstructed computationally to generate a 3D volume or tomogram. The resulting tomogram shows the interior of the cell in three dimensions at the highest achievable resolution without fixation or perturbation (see Figure 1). Detailed workflows for cryo-FIB milling and electron tomography (cryo-FIB-ET) are available (82, 123, 143).

ADVANCES IN THE CRYO-FOCUSED ION BEAM WORKFLOW

Recent improvements in the cryo-FIB-ET workflow have dramatically increased sample throughput, experimental handling, and improved data collection. For adherent cells, the first step in this workflow is seeding cells on an EM grid. Adaptation of micropatterning to EM grids enabled the control of the position, size, and shape of adherent cells on the EM grid (138). Added to this, 3D-printed holders for EM grids can minimize grid handling mishaps (49). It is unfortunate, after preparing a lamella, to discover that it is not suitable for high-resolution TEM data acquisition. This unsuitability could be due to lamellae being too thick, broken, badly vitrified, or contaminated during transfers, among others. Recent advances in the FIB milling process have resulted in more optimal lamellae that are more stable, have reduced ice contamination, and notably can be generated through automation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The stability of the lamella has been made more robust through stress-reducing expansion joints, or trenches, around the lamella site (157). A cryo-glove box was developed to reduce contamination during sample transfer (136). Incubating samples in 10% glycerol improved the vitrification of large samples, i.e., the peripheral nervous system tissues from *Drosophila* (7). Another challenge is that milling can result in curtains due to differential sputtering rates in cellular compositions, ice contamination, or milling at high currents for an extensive period of time (82, 123). Platinum deposition and sputtering alleviate some curtaining effects, as described below. To alleviate persistent curtaining and improve visual interpretability, curtain artifacts can be removed in silico, with regions of interest targeted on the postprocessed image (7). Finally, through FIB milling automation, sample throughput has improved, and a user is now able to generate up to 20 lamellae in a session through software automation (79, 80, 136, 163).

Newer methods for dealing with a diverse set of samples have expanded what can be imaged by this technique. Tissues, or other thick specimens, cannot be fully vitrified by plunge freezing due to inadequate heat transfer (133) but can be vitrified through HPF. After HPF, a section of the specimen is carved using the FIB, lifted out from the sample, transferred to a new grid placed in the same shuttle within the FIB-SEM instrument, and then further thinned down to generate a 100–300 nm lamella through a method referred to as lift out (93). A recent study used in situ sputtering from the micromanipulator to enable cryogenic preparation of specimens for atom probe tomography by FIB, enabling the analysis of silicon content within a sample (39).

As an alternative to lift out, the inventive waffle method makes cryo-FIB-ET more amenable to thick specimens; specimens that orient preferably on an EM grid, e.g., bacteria that lie along their long axis; and specimens that require a high local concentration, e.g., cells that need to grow to confluency (75). In this method, the sample is applied to the backside of an EM grid and is then high-pressure frozen, thereby achieving high uniform density, which facilitates cells orienting in multiple directions (75).

Recently, the types of ion sources used during the cryo-FIB milling process have been expanded to include plasma from gases including oxygen, nitrogen, argon, and xenon. Initial studies suggest that these gases may induce less damage from the ion beam and lead to higher-resolution data (12, 129).

Data Collection

After producing a beautiful (or at least usable) lamella, the user must decide on a data collection scheme (Figure 2). A tilt series is a collection of images of a sample at different tilt angles. Each of these images is called a projection, as it collapses the 3D information of the sample along the optical path of the microscope into two dimensions (58). Because biological macromolecules are low atomic weight, they are considered to be weak-phase objects, as the generated contrast is weak. Image contrast is achieved by shifting the focal plane via defocusing the objective lens (62). This image defocus, along with other factors, introduces variations in the image described by the contrast transfer function (CTF) that can be corrected to retrieve a more accurate representation of the sample (51). Tilt series acquisition that produces a high-accuracy reconstruction in three dimensions relies

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

on several factors, including stage and sample stability; lamella thickness; location of the feature of interest within the lamella; and, once the data are acquired, the ability to align each projection in the tilt series and CTF correct it. During data collection, additional images are typically acquired to ensure that the field of view remains centered in the area of interest (*XY* tracking) and at the same defocus (*Z*). Subsequently, to align the tilt series accurately for 3D reconstruction, features in the images of the tilt series must be followed from one image to the next (fiducials). Gold beads are electron dense and are utilized extensively for in vitro cryo-ET; these fiducials enable accurate focus and tracking during collection and tilt series alignment post-collection (147). In most cases, membranes and other features present in the cells can be used as fiducials for tilt series alignment. In one attempt to improve alignments for in situ studies, researchers added gold beads to the cell culture media; these gold beads were endocytosed, becoming fiducials within the lamella (13).

How the tilt series should be collected depends on several factors, including the scientific question, how the data will be analyzed, and sample quality. Software packages offer flexibility in tilt series acquisition; these include the academic packages SerialEM and Leginon (98, 134) and the commercial packages Tomography 5 (Thermo Fisher Scientific), Latitude (Gatan), and Recorder (JEOL). These packages allow the user to control parameters such as acquisition scheme, tilt range, tilt increment, dose, magnification, and target defocus. SerialEM is a broadly used and reliable package that is ideal to test new schemes while offering out-of-the box solutions, including automated tomography data collection (98). National Institutes of Health funding and user-contributed support of SerialEM facilitates such innovation by providing a platform on which users can develop and disseminate their own schemes for data collection.

A major consideration when acquiring data is the accumulated radiation damage to target areas of acquisition (65). When collecting 2D data [as in single-particle analysis (SPA) and 2D template matching, introduced below)], a single movie is acquired with a total fluence (commonly referred to as dose) of 60–120 e^{-}/A^{2}. In tomography, the total dose that a sample experiences has to be distributed among many images in the tilt series. The choice of how to distribute the dose determines how well the individual images can be aligned to others in the tilt series and which images have the most information: The projections generated from early stages in data collection will contain the least amount of radiation damage and the best-quality high-resolution information. A variety of tilt series acquisition schemes are available depending on such considerations.

Collection schemes include continuous collection, in which the sample is tilted to the maximum tilt angle on one side (e.g., –70°) and then toward the maximum tilt angle on the other side (e.g., +70°). The benefits of this scheme are speed of data acquisition and minimal stage perturbations. However, when acquiring a tilt series, the effective thickness of the sample increases as the sample tilts from 0° to 70° because there is more material for the beam to pass through at high tilts (the thickness increases as 1/cos of the tilt angle). Thus, the detriment of continuous collection is that the initial dose, when the sample is most undisturbed by radiation, is spent on high-tilt images, where the sample will be the thickest, resulting in a lower-contrast image. A commonly used bidirectional collection scheme starts at 0° (or at the angle where the lamella is thinnest, compensating for its pretilt), moves

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

toward the highest angle in one direction, and then returns to the initial position, collecting until it reaches the maximum tilt angle in the other direction. A bidirectional tilt series will result in two halves, and since the adjacent projection images at the lower tilts will have experienced different amounts of accumulated dose, there is often a problem aligning the tilt series. Additionally, since the sample is thinnest at the low tilt angles, it is at these angles that the images can be attained at high resolution; however, in the bidirectional scheme, at low tilt angles on one side, the sample has received significant electron dose, and thus high-resolution information is lost (139). To maximize initial dose at lower tilts while still producing a smooth reconstruction, Hagen et al. (67) developed the dose-symmetric (DS) scheme, also known as the Hagen scheme, in which acquisition begins at 0° (or the lamella pretilt) and alternates between positive or negative tilt angles along the tilt range. A thorough comparison among these schemes via STA of in vitro data concluded that the DS scheme resulted in the highest attainable resolution (139).

Beyond optimizing for individual tilt series, new collection schemes have been described that can (*a*) decrease the amount of time for sets of tilt series to be acquired and further boost the throughput of data collection, (*b*) improve the CTF estimation and correction, and/or (*c*) collect data in a way that allows better utilization of high-resolution information (Figure 2).

Single versus dual tilt and continuous acquisition.—Tomography data typically suffer from a missing wedge, since samples cannot be tilted to obtain tilt projections in the full range (–90 to 90°) (124). Multi-tilt axis tomography can reduce this missing information from a wedge to a pyramid or cone (112). The Titan Krios (Thermo Fisher Scientific) was designed initially to tilt along two orthogonal tilt axes. However, it was determined that, due to radiation damage, dual-axis tilt series collection did not offer an advantage over single-axis tilt series collection (28, 41). Thus, new-generation microscopes have a single-tilt holder with improved mechanical stability. Using a single-tilt holder, researchers demonstrated that one could dramatically reduce data collection time to minutes by setting the camera to acquire one continuous movie during the tilt scheme (28, 41), with the electron beam blanked when the stage moves between tilt angles and for a preset amount of time afterward to allow the stage to settle; this is referred to as the fast-incremental method (28, 41).

Beam shifts reduce acquisition time.—The rate-limiting step for TEM data collection is waiting for minimal stage drift following stage movements due to stage motor vibrations (27). In SPA, it is routine to use beam shifts rather than stage shifts, obviating the need for settling times and thus increasing the number of attainable targets, as a central focus and tracking area is employed and the beam is then tilted to desired target sites for data collection (27). Data collection is >720 micrographs per hour for a K3 using image shift, a roughly 10-fold increase compared to stage shift (111). Post-collection, the beam tilt is estimated and corrected during CTF estimation (167).

The challenge to implementing this method for tomography is that the distance between a central focus and tracking area and the target point changes during the tilt series acquisition. Beam-image shift electron cryo-tomography (BISECT) was able to adapt this beam-image shift for in vitro tomography by determining relevant geometric constraints (16). BISECT

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

reduced tilt series acquisition time to mere minutes, without compromising data quality, when comparing STA results to collection by the DS method (16). In multishot tomography, researchers implemented beam-tilt image shift by acquiring a multishot tilt series along the tilt axis (77). Unfortunately, cellular features of interest within a lamella are not likely to be arranged neatly along the tilt axis. To overcome this problem, a third group developed the parallel cryo-electron tomography (Pace-Tomo) acquisition scheme, which utilizes a geometric model of the lamella that accounts for the lamella pretilt and sample geometry, and also developed an empirically determined defocus ramp that reduces errors in the tilt axis offset (42). With both Pace-Tomo and multishot tomography, researchers achieved 8 Å reconstructions of an in situ ribosome (16, 42).

The best of both worlds? Hybrid subtomogram analysis and single-particle analysis.—In tilt series collection, a total electron dose of 80–200 e^{-}/A^{2} is divided along the entire tilt range. This results in a low SNR for each tilt image, since 1–3 e^{-}/A^{2} is spent on a tilt image, depending on total dose, tilt increment, and tilt range. A few hybrid STA-SPA (hSTA) workflows have been described wherein a higher-electron-dose image is acquired at the 0° tilt angle, followed by a DS tilt series for the remaining projections (121). The initial high-dose image at 0° has a higher SNR, resulting in a better defocus estimation and CTF correction. Using hSTA, Sanchez et al. (121) improved the resolution of the Tobacco Mosaic Virus (TMV) from 7.2 Å to 4.4 Å compared to a DS tilt series. Similarly, TYGRESS implemented a hybrid workflow: At 0°, the specimen is exposed to an initial high dose (30 e^{-}/A^{2}), followed by a regular low dose; a DS tilt series is collected afterward (131). In subsequent STA, the high-dose image is used for CTF estimation.

Capturing the full picture (montage data collection).—The choice of magnification, and thus field of view, is a compromise between the area that needs to be included in the tilt series and the pixel size, which determines the limit to resolution (151). Another consideration is that higher magnification also results in higher total doses: The dose in an individual image increases as $1/r^{2}$, where r is the pixel size. For research inquiries that would benefit from a large field of view without decreasing magnification or resolution, it would be advantageous to obtain one exhaustive tomogram of the entire lamella. This idea was initially developed for tomography of data collected from resin-embedded samples (113). By collecting cryo-ET data composed of overlapping fields of view, two studies tiled multiple images to create one massive montaged tomogram with a large field of view (110, 161).

2D template matching.—For in situ STA, the ability to obtain high-resolution structures, or to unveil multiple conformational and compositional states, can be limited by the number of particles that can be extracted from the data. This limitation can be due to abundance of the macromolecule of interest in cells, the number of in situ conformational or compositional states (e.g., transient interactions), or other factors. Even more challenging is how to find these particles in the crowded environment of the cell, where most proteins have similar shapes and molecular weights. How can we identify our particles of interest and generate high-resolution reconstructions? One powerful idea is to avoid tilt series acquisition and instead acquire a low-defocus and low-dose (approximately $30e^{-}/A^{2}$) 2D projection (120). Using the high-resolution information, users search for particles by template matching

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in two dimensions, determining the position in X-Y and orientation within the sample as defocus estimation is used to determine the localization along the optical axis (Z). Once these parameters are determined, the particles in two dimensions can be used for averaging to obtain high-resolution 3D reconstructions (89). In a follow-up study, Lucas et al. (90) used this method to locate ribosomal subunits in the nucleus of a yeast cell and found that homolog templates were enough to find relevant particles, with which they could identify and resolve 60S maturation intermediates in situ. Expanding this idea to the entire lamella, in the defocus correction large area cryo-EM (DeCo-LACE) method, a large field of view was collected at low defocus values, enabling the identification of ribosomes at low defocus within the cell by 2D template matching (44). This powerful technique promises to be the key to locating proteins in crowded environments and using high-resolution information to determine their structures, provided that there are templates to search for and that enough particles can be found within the data (see Figure 3).

Phase plate.—In TEM, contrast is generated by defocusing the objective lens, and spatial information is recovered by estimating and correcting the CTF. The Volta phase plate (VPP) generates phase contrast by introducing a phase shift between the scattered and unscattered electron waves (35). Since the VPP is a physical material that is exposed to the EM beam, the phase shift evolves over time and needs to be monitored (35). Examples of VPP use for enhancing contrast for tomographic studies include initial work visualizing the nuclear periphery by cryo-FIB-ET (92) and structural analysis on in situ chromatin (23, 50). The VPP does not reliably outcompete standard methods for high-resolution data determination (139). However, it has excited the field due to the possibilities of phase contrast imaging in TEM. It was recently demonstrated that the laser phase plate, a high-intensity continuous-wave laser beam, is capable of generating phase contrast imaging in TEM (128). Because the phase shift is generated by a laser and not a material, continuous exposure to the TEM beam would not affect data acquisition over time.

Scanning transmission electron microscopy.—What if we were able to acquire more than just projection images, i.e., more than just a series of gray values from our biological samples? Pioneering work using cryo-scanning TEM (STEM) has demonstrated that it is feasible to obtain spectroscopic information from biological samples (156). STEM uses a weakly convergent beam at high energy and generates an electron diffraction pattern that is then detected by bright field, annular dark field, or high-angle annular dark field detectors (43). As a proof of principle, STEM with integrated differential phase contrast mode resolved TMV to 3.5 Å but imaged with only 30 $e^{-}/Å^{2}$ (85). Such low doses would be ideal for in situ tomography.

Data Reconstruction and Analysis

Interpreting a tomogram is challenging because of the low SNR; recall that biomolecules are weak-phase objects, composed of elements of similar atomic masses, and are radiation sensitive. Additionally, biomolecules exist in a very crowded medium, and many of them have very similar overall shapes. Unequivocally identifying individual proteins or biomolecular complexes is challenging, and the type of analysis that must be performed depends on the focus of the study. Cryo-ET can be used as an imaging method or driven by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

structure determination, whereas STA is used to determine the architecture of biomolecular complexes in situ. Computational algorithms improve the interpretability of a reconstructed 3D tomogram, facilitating data annotation (segmentation) and particle identification for STA.

Reconstruction methods.—After tilt series acquisition, tilt images are aligned and then reconstructed to generate a 3D volume or tomogram; typical methods include weighted back project (WBP) (118), simultaneous iterative reconstruction technique (SIRT), or simultaneous algebraic reconstruction technique (SART) in Image Modeler (IMOD) (100). In WBP, the 2D projection images are back-projected with a weighting function that preserves the high-frequency information that is necessary for further structure determination from the resulting 3D tomogram (57).

The quality of the reconstruction will depend on the accuracy of the tilt series alignment and the accurate correction of local motion. In situ cryo-FIB-ET cannot easily utilize fiducials to solve the tilt series alignment, as with in vitro tomography; thus, recent solutions for image misalignment work computationally. Warp and AreTomo (Alignment and Reconstruction for Electron Tomography) (165) solve for and correct for in-plane rotations, translations, and local motion. While Warp utilizes individual particles averaged from a tomogram to improve the tilt series geometry, AreTomo solves these parameters and applies those solutions to the whole tomogram.

Denoising filters, deconvolution, machine learning (ML), and deep learning (DL) algorithms improve the interpretability of noisy cryo-EM data by acting on either the input 2D projections, during the reconstruction process, or the 3D reconstruction. Denoising filters enhance the SNR and visual interpretability of whole tomograms; these filters include the non-local means filter, median filter, Sigma filter, and Gaussian filters (55). SIRT and SART filters enhance low-frequency information that is suitable for visual interpretation (147). Additional post-reconstruction filters include the non-linear anisotropic diffusion filter, which smooths and enhances edges (56).

ML and DL algorithms, which were developed for restoration of any type of image, have been adopted by denoising programs specifically for EM data. ML strategies such as Noise2Noise, Noise2Void, and Noise2Self are particularly useful when the ground truth is unknown—as it is for cryo-EM data (11, 18, 86, 137). Programs that denoise TEM projection images or tomographic volumes through DL include Warp, CryoCARE, Topaz-Denoise, and IsoNet (11, 19, 87).

Deconvolution for missing wedge mitigation.—Due to the inherent geometry of a lamella slab and data acquisition, there is a missing wedge of information in the *Z*-direction, and data collection is limited to +/–60° due to the sample holder tilt range (124). The inability to correct for this missing information, or to account for its absence, can cause artifacts in downstream analysis. Deconvolution is regularly applied to fluorescence microscopy images (122) and, more recently, STEM cryotomography (150). Croxford et al. (31) found that applying an entropy-regularized deconvolution algorithm to cryo-ET data improved the reconstruction by modeling the point spread function (PSF) of the TEM beam

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in three dimensions, then deconvolving the back-projected 3D image with the 3D PSF to fill the missing wedge, resulting in a less isotropic reconstruction. Using experimental and synthetic data, they demonstrated that their deconvolution method fills in information in the missing wedge lost during either tilt series acquisition or WBP.

Missing wedge mitigation is also performed via isotropic reconstruction of electron tomograms with DL by IsoNet (88). IsoNet adds Gaussian noise to a training set and then iteratively improves the SNR by attempting to eliminate the added noise through DL (88). The information learned during training is then used to fill the missing wedge (88). Standards for denoising tomograms by DL would be helpful as this field progresses, as would guidelines for appropriate user implementation during data analysis. Conservative uses of a tomogram denoised through ML include visual interpretation and particle picking, always referring to the original data for further STA.

Data annotation.—After improving the contrast and SNR of a tomogram, annotation enhances the interpretability by defining the relevant features. The process of annotating features in a tomogram is referred to as segmentation. Membranes and filaments are continuous 2D and 1D objects that are readily identifiable in tomograms. Manual membrane segmentation of cellular organelles was initially traced by hand, a time-consuming process. Algorithms that automate at least part of this process include TomosegmemTV (95), Membranorama (153), and the Surface Morphometrics toolkit (5). A growing number of programs identify and annotate cellular features through artificial intelligence: DeepFinder identifies macromolecules (104), and MemBrain identifies membranes (83), as well as programs for filament identification (37).Template-free detection and classification of membrane-bound complexes in a tomogram is another useful tool (96). Complete annotation of a tomogram (membranes, filaments, and macromolecules) can be performed in EMAN2 (26). Software to visualize these annotations and utilize embedded analysis tools include Amira and DragonFly (114, 132).

Subtomogram Analysis

There are several situations in which in situ tomography is an ideally suited technique, including when the goal is to determine the structure of macromolecular assemblies as they exist in their native environment; when the proteins are deeply embedded in their environments and cannot be faithfully reconstituted outside of cells (9, 108, 130, 158); when the assemblies are transient, or multiple compositional or conformational states exist (1, 2); and when contextual information will never be faithfully recapitulated in vitro (e.g., chromatin topology). The molecular architecture of in situ complexes can be resolved by STA and combined with integrative modeling, facilitating the proposal of new structures and models. STA typically entails multiparticle refinement and its predecessor, subtomogram averaging.

Subtomogram averaging.—Following tomogram generation and the visual enhancement of choice, if your molecule of interest is present, then you would explore STA methods. We refer readers to more detailed reviews that describe current STA methodology in theory and in practice (116, 117, 147, 164). First, frames are aligned to account for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

beam-induced motion (BIM). Then, the CTFs of the tilt images are estimated and corrected; however, because tilted images have a defocus gradient, CTF estimation correction must occur in three dimensions in its local position to accurately determine the CTF for a given particle (140). The tilt series are aligned, and a tomogram is reconstructed (99). The next step is particle picking or template matching for molecules of interest, after which subtomograms around that particle are extracted, and those subtomograms are then aligned to a reference (17). Software packages for subtomogram averaging include Dynamo, EMAN2, emClarity, RELION, and StopGap (10, 14, 24, 148). Structures determined using STA on in situ data are being published at an ever increasing pace. Often, the structures obtained test or generate existing hypotheses that can only be addressed in their natural environment. Examples include the nuclear pore complex (1, 9, 105, 166), bacterial and eukaryotic ribosomes (42, 47, 84, 107, 137), microtubules (54, 64, 135), proteasomes (2, 4), LRRK2 bound to microtubules (149), the Arp2/3 complex (48), the COPI coat (22), nucleosomes (23), a bacterial chemosensory array (20), and a bacterial gap junction (152).

Multiparticle refinement.—In situ structure determination improved significantly with the implementation of a framework in which relationships between particles within the same tomogram are solved as part of a larger system. Due to BIM and radiation damage, particles appear to move in different trajectories through a tilt series. By determining the trajectories of one set of particles, one can facilitate the alignment and translation of neighboring particles within that system (6, 137). Implementation of multiparticle refinement yielded a 3.4 Å reconstruction of an in situ stalled ribosome (137). Initially referred to and described as constrained single-particle tomography, multiparticle refinement was fully realized when it was implemented through emClarity and Warp/M (6, 69, 137). By solving the particle trajectories of one species, one can improve the trajectories of another; for example, researchers performed STA on bacterial ribosomes and then used the optimized tilt series to improve the trajectories of the target, a jumbo phage nuclear shell, chimallin, thereby improving the final reconstruction (84).

A major source of difficulty in implementing STA is migrating data between software programs. It is nontrivial to properly convert particle coordinates, metadata, etc. (21). It is not uncommon to jump among Dynamo, Warp/M, and RELION; for instance, the in-cell structure of Arp2/3 relied on all of these software packages (48). A collection of scripts that makes the move among Dynamo, Warp/M, and RELION seamless was recently published, a win for all Team Tomo members (21). Additional scripts for converting RELION-v4 starfiles are available via the RELION-4 Github.

Making the most out of angular information.—A very powerful algorithm, NEMO-TOC (NEighboring MOlecule TOpology Clustering), utilizes the coordinates and the Euler angles determined from STA to facilitate the clustering and interpretation of neighboring macromolecules such as polysomes along the same messenger RNA chain (74). There is tremendous potential to apply fully utilized refined particle coordinates to map structures and orientations in their native environment. For instance, future applications using the coordinates and refined Euler angles of in situ nucleosomes would advance our understanding of chromatin chain topology directly inside the cell.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

QUANTITATIVE CELL BIOLOGY AND BIOPHYSICAL MODELING

A cellular tomogram, not unlike a painting by a Dutch master, is rich in both detail and contextual information, with diverse equally mesmerizing vignettes (T. Laughlin, personal communication). Where should one begin the analysis? One could solve the ratio of active to inactive proteasomes within a neuron (2) or measure exactly the differences in dimensions and curvature between two organelles (5, 91). Additionally, since bacteria are comparable in size to the light diffraction limit (76), their internal structure and modeling are challenging for traditional light microscopy methods, and thus many unexpected discoveries and molecular mechanisms have been resolved using cryo-FIB-ET (25).

This section highlights a few examples in which in situ tomography data have been used to better model the characters in a cell and how they interact in a quantitative and biophysical manner. This list is by no means exhaustive, but it serves as an illustration of the power of the technology as a quantitative imaging tool. The earliest work performing in situ cryo-ET studies was on the thin cellular regions, which are easily vitrified and thin enough (<300 nm) that the TEM beam could penetrate (9). Later,visualization of actin comet tails in cells infected by the bacterium *Listeria monocytogenes* facilitated biophysical models for actin branching based on X, Y, and Z coordinates determined in their native state (73).Recent work that beautifully captured actin-mediated endocytosis directly provided the basis for biophysical modeling of this fundamental event (130).

Moving further into the cell, cryo-FIB-ET studies have illuminated junctions between organelles where membrane contact sites communicate and exchange contents. Membrane contact sites are sites of lipid transfer, approximately 20 nm structures, and are difficult to resolve by other techniques. The endoplasmic reticulum (ER) contacts many other organelles, as it is the lipid factory responsible for synthesizing structural phospholipids, sterols, and storage lipids (115). What is the molecular basis for lipid exchange? Direct visualization of the molecules between the plasma membrane and the cortical ER (cER), followed by correlative light microscopy EM (CLEM), made it clear that extended synaptotagmins (tricalbins in yeast) form 16–24 nm bridges between these two organelles at regions of high cER curvature and that these bridges are Ca^{2+} dependent (30, 71).

Through STA and integrative modeling, Wozny et al. (158) recently proposed a model for lipid transport for the ER–mitochondrial encounter structure (ERMES) in which three ERMES synaptotagmin-like mitochondrial lipid-binding protein (SMP) domains stack to form a continuous structure or supramolecular organization that facilitates lipid transport between the ER and mitochondria (158).

Cryo-FIB-ET preserves membrane features extremely well and allows researchers to better visualize membrane remodeling events that occur during stress, starvation, and apoptotic conditions. Stress and starvation elicit the de novo formation of an entirely new organelle, the autophagosome. How the autophagosome is nucleated, which membranes supply lipids to the growing autophagosome, and how subsequent fusion with the lysosome or vacuole occurs are open questions. Visualizing and quantifying the entire process in nitrogen-deprived yeast via cryo-FIB-ET led to a few striking conclusions. It was revealed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that general macroautophagy does not require cargo targeting. Through nearest-neighbor analysis, it was determined that the phagophore comes in very close contact with many membranes—lipid droplets, the ER, the nucleus, mitochondria, and the vacuole—potentially explaining conflicting reports about which organelles supply the lipids (15). Finally, it was found that the curvature of the phagophore rim changes shape during phagophore growth, which is perhaps related to how the phagophore rim expands (15).

Observations that Yield More Questions

Cryo-FIB-ET has discovered new biology (25) and likely will continue to do so, as it is currently the highest-resolution imaging method for cell biology capable of capturing intact molecular networks, and because the imaging source is agnostic to the sample, all molecules are observed. In this section, we highlight a few unexpected observations made by cryo-FIB-ET that have raised additional questions.

While investigating the *Plasmodium* life cycle, researchers discovered microtubule protofilament diversity: 13–18 protofilament microtubules, as well as doublets, triplets, and quadruplets in the gametocyte (52). They found a pseudohelical protein within the microtubule lumen of the sporozite, the *Plasmodium* motile spore-like stage; its identity remains unknown (52). De novo structure determination or AlphaFold mining of the *Plasmodium* genome with structural constraints could perhaps yield its identity, thus realizing visual proteomics.

Models of vesicular transport through the Golgi (from *cis*-Golgi, to Golgi stacks, to *trans*-Golgi) have consolidated around the vesicle maturation model. A rather striking STA result determined that there is an array of proteins in the narrow *trans*-Golgi lumen of *Chlamydomonas* (46). These structures show high periodicity, with interactions between *cis*- and *trans*-facing arrays arranged in a zipper-like fashion, and potentially unidentified glycosyltransferases (46). As these proteins occupy the central portion of the Golgi, it suggests that modified cargo is forced into the Golgi periphery by locally narrowing the *trans*-Golgi lumen (46).

Cell biology studies on how the mitochondria undergoes fission implicated the ER in this process, as it appeared from live-cell confocal fluorescent microscopy that the ER encircled the mitochondria (60). However, in situ cryo-FIB-ET showed that mitochondrial fission directly involved actin and septin right before fission (91). Whether the ER could become involved, after actin and septin, and encircle a mitochondria is currently an open question.

How to Find Your Favorite Macromolecule

Unequivocally identifying a molecule of interest within in situ cryo-ET data remains one of the largest challenges in using this method. When peering at your tomogram, your eyes are drawn to recognizable features: large macromolecular assemblies, lipid bilayers, filaments. Thus, we have a problem—smaller, less recognizable macromolecules are hiding in plain sight, as it is hard to unequivocally identify a molecule only by a tomogram. Tools to identify your molecule of interest include cryo-fluorescence microscopy (cryo-FM), metal-based tags, and molecular tags. This section covers these technologies and what caveats and challenges exist for each.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CLEM

There are two steps in which fluorescence microscopy is extremely useful in targeting and identifying molecules of interest. CLEM is a well-established tool for correlating fluorescence images and EM data. Fluorescence images taken under cryogenic conditions (cryo-FM) can be used to identify which cells among the ones in the grid contain the feature of interest to generate a lamella, for instance, when thereis heterogenous protein expression, as with a transient transfection; stochastic expression due to lentiviral integration; or a rare cellular event. The resolution of a cryo-FM image is roughly 250 nm in X-Y and 500 nm in Z (36). As the desired lamella thickness is 100–300 nm, it is possible to mill through your target if you are relying on fluorescence images limited to 500 nm. The second CLEM step overlays the fluorescence image on a low-magnification TEM image of the lamella to identify molecules or features of interest. Unfortunately, correlating protein localizing between fluorescence microscopy and the TEM image is challenging, as it requires correlation between features in both, and is thus best applied to membrane-bound targets, microtubule-bound targets, or discrete assemblies rather than dispersed cytosolic targets. Nevertheless, many studies have been made possible only through CLEM. The recent development of CLEM systems installed inside the dual beam FIB-SEM microscope (63) promises a higher yield and throughput of lamellae containing the features of interest, and commercial devices using these systems are now available, including Meteor (Delmic) and iFLM (Thermo Fisher Scientific).

What about superresolution microscopy? Ideally, we would want to locate proteins of interest within a tomogram with single-molecule precision. The challenges with adopting superresolution microscopy methods for cryogenic conditions include sample devitrification from the laser source, the poor numerical aperture of an air objective lens (36), the difficulty of photoactivating a fluorophore at cryogenic conditions, and correlating the TEM and fluorescence data to high precision (125). A proof-of-principle superresolution experiment using single-molecule active control microscopy (SMACM) utilized a photoactivable protein (PAmKate) to perform correlative cryogenic SMACM and cryo-ET (CIASM, pronounced like chasm) to identify the localization of caulobacter proteins (34). One of the main challenges in this work involves heating (and devitrifying) the sample with the laser, which limits the number of fluorophores that can be detected. Using different substrates, Dahlberg et al. (33) managed to increase the density of localizations and improve correlative imaging. Use of this imaging modality in lamellae remains to be implemented, as thermal conductivity in frozen cellular samples is low. Another exciting development in the attempt to fully leverage superresolution capabilities was achieved through a custom-built liquid helium–cooled microscope, which images samples under vacuum conditions. These high-pressure frozen specimens were imaged by cryo-superresolution microscopy, thereby facilitating the study of ultrastructure relationships in intranuclear vesicles, the ER, and chromatin transcriptional domains between two imaging states: superresolution fluorescence and electron microscopy (70).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Other Molecular Tags

Besides tagging proteins of interest with fluorescent proteins, use of other types of tags that can be visualized directly in the TEM would dramatically expand the amount of the proteome that is accessible to cryo-ET. One possibility is to use signpost origami tags (SPOTs), which recognize and bind green fluorescent protein (GFP) through an aptamer composed of DNA origami (68). The effective distance from the target to the SPOT is roughly 70 nm, which may or may not be achievable and may disrupt the molecular networks, depending on the query.

Using nanoparticles composed of atoms with high atomic numbers to find your molecule of interest would be ideal, as biological material is mostly made of the lighter elements—carbon, nitrogen, oxygen, phosphorus, trace metals, etc. Heavy elements such as gold or iron would appear as a darker contrasting object, as the atom's nucleus is more dense by comparison to the biological sample and would scatter more electrons from the imaging source. An example of this is the FerriTag system, in which the molecule of interest is expressed as a fusion construct to GFP-FKBP, and dimerization occurs when rapamycin is supplied and FKBP binds to a coexpressed fusion of an FRB-mCherry-ferritin cage (29).The large ferritin cage is roughly 40 nm away from the molecule of interest (29).Another ferritin-based labeling strategy expressed a ferritin cage within *Escherichia coli* and was used to target specific proteins of interest (146). Additionally, a 2.2 nm nanogold particle was conjugated to designed-ankyrin repeat proteins (DARPins), which were engineered to recognize extracellular integrins on a mammalian cell and visualized by cryo-ET (32).

CRYO-ELECTRON TOMOGRAPHY BY THE NUMBERS

Above, we describe technical hurdles in cryo-ET and relevant solutions that allow us to capture macromolecules in near-native conditions; however, a greater challenge is selecting or designing the right model system. The scientific questions driving a project will determine how the data are acquired, what kind of analysis is performed, and what amount of data is required. If structure determination is the goal, then what is the resolution required? How abundant is the complex or event of interest in the cell, and thus, what is the likelihood of capturing it in a tomogram? If it is present, can it be unequivocally identified? What is the potential conformational and compositional variability of the protein or complex? The bottom-line question is how many instances of an event must be captured to better understand its structure–function relationship in situ? From this question follows another: Is in situ absolutely required?

Let us consider the fact that a typical mammalian cell is roughly 2,000 μm^3 (103) (BioNumber Identification 108244). A typical lamella is a 3D slab that is ideally approximately 100–200 nm in thickness and covers an area of approximately 10–20 μm per side (approximately 100–400 μm^2); thus, the lamella volume is approximately 10–80 μm^3, covering 0.5–4% of the cell. With the magnifications typically used for cryo-ET, which result in an imaging area of 0.4 to 2 μm per side, a tomogram covers significantly less than a percentage of the total cell volume. Knowing the total copy number (or the local concentration) of your protein of interest is necessary to obtain an estimation of how many lamellae, tomograms, and particles will be necessary to observe your chosen macromolecule

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and perform relevant analysis. To answer the question of how many lamellae and tomograms are needed to answer your question or solve a structure, estimate how many copies one could reasonably achieve per tomogram, how heterogeneous these instances are expected to be, and what resolution is necessary to answer the question. The size and cellular abundance of a given macromolecule (or its local concentration) are relevant when estimating how many lamellae, tomograms, and particles you will need to perform STA. Not surprisingly, the more abundant protein complexes (ribosomes, nucleosomes) will have a greater ratio of particles per tomogram compared to less abundant complexes (the nuclear pore complex). However, since the nuclear pore complex and the ribosome are megadalton-sized complexes, they can be easily identified in a crowded tomogram, and their large size also facilitates particle alignment during STA. Such back-of-the-envelope calculations are useful when estimating how much data (roughly) are required to answer a given question (see Figure 3).

THE STRUCTURAL REVOLUTION

Before the resolution revolution in SPA (81), integrative modeling was used to bring together data from various sources (X-ray, cryo-EM, mass spectroscopy, cross-linking, proteomics, and others) to model the architecture of macromolecular complexes. This tradition has continued as cryo-ET has begun to produce maps of challenging complexes in situ (1, 149). More recently, the advent of AlphaFold has potentialized this methodology: Models from AlphaFold can be used to interpret the cryo-ET data, e.g., using template matching, and to assist in building molecular models from in situ maps obtained through STA (53, 106).

OUTLOOK

Cryo-ET has rapidly established itself as a powerful technology to bring structure to cell biology. Although lacking in temporal resolution, this technique is currently the highest-spatial-resolution imaging tool available for cell biology, and it can be used to make new discoveries or to generate new hypotheses that can be further tested with other methods such as live cell imaging and proteomics. It can also produce novel structures of macromolecular complexes embedded in their environment or perform censuses on the number and composition or conformation of molecular complexes in cells, as well as other quantitative analysis. Future advances and automation will democratize this technology and allow it to become a widely available and powerful tool for the cell biologist.

ACKNOWLEDGMENTS

We thank Dorotea Fracchiolla and Tara Bensen for illustrating the cryo-FIB-ET workflow in Figure 1. We thank members of the Villa lab for their helpful comments on the manuscript. L.N.Y. is supported by a National Institutes of Health NCI K00 award (5K00CA223029-05). E.V. is a Howard Hughes Medical Institute Investigator.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

LITERATURE CITED

1. Akey CW, Singh D, Ouch C, Echeverria I, Nudelman I, et al. 2022. Comprehensive structure and functional adaptations of the yeast nuclear pore complex. Cell 185(2):361–78.e25 [PubMed: 34982960]
2. Albert S, Schaffer M, Beck F, Mosalaganti S, Asano S, et al. 2017. Proteasomes tether to two distinctsites at the nuclear pore complex. PNAS 114(52):13726–31 [PubMed: 29229809]
3. Allegretti M, Zimmerli CE, Rantos V, Wilfling F, Ronchi P, et al. 2020. In-cell architecture of the nuclear pore and snapshots of its turnover. Nature 586(7831):796–800 [PubMed: 32879490]
4. Asano S, Fukuda Y, Beck F, Aufderheide A, Förster F, et al. 2015. Proteasomes. A molecular census of 26S proteasomes in intact neurons. Science 347(6220):439–42 [PubMed: 25613890]
5. Barad BA, Medina M, Fuentes D, Wiseman RL, Grotjahn DA. 2022. A surface morphometrics toolkit to quantify organellar membrane ultrastructure using cryo-electron tomography. bioRxiv 2022.01.23.477440. 10.1101/2022.01.23.477440
6. Bartesaghi A, Lecumberry F, Sapiro G, Subramaniam S. 2012. Protein secondary structure determination by constrained single-particle cryo-electron tomography. Structure 20(12):2003–13 [PubMed: 23217682]
7. Bäuerlein FJB, Pastor-Pareja JC, Fernández-Busnadiego R. 2021. Cryo-electron tomography of native Drosophila tissues vitrified by plunge freezing. bioRxiv 2021.04.14.437159. 10.1101/2021.04.14.437159
8. Baumeister W 2022. Cryo-electron tomography: a long journey to the inner space of cells. Cell 185(15):2649–52 [PubMed: 35868271]
9. Beck M, Förster F, Ecke M, Plitzko JM, Melchior F, et al. 2004. Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. Science 306(5700):1387–90 [PubMed: 15514115]
10. Bell JM, Chen M, Fluty A, Ludtke SJ. 2018. A complete workflow for cellular tomography and subtomogram averaging in EMAN2. Microsc. Microanal 24(S1):866–67
11. Bepler T, Kelley K, Noble AJ, Berger B. 2020. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. Nat. Commun 11:5208 [PubMed: 33060581]
12. Berger C, Dumoux M, Glen T, Yee NB- Y, Mitchels JM, et al. 2022. Plasma FIB milling for the determination of structures in situ. bioRxiv 2022.08.01.502333. 10.1101/2022.08.01.502333
13. Berger C, Ravelli RBG, López-Iglesias C, Peters PJ. 2021. Endocytosed nanogold fiducials for improved in-situ cryo-electron tomography tilt-series alignment. J. Struct. Biol 213(1):107698 [PubMed: 33545353]
14. Bharat TAM, Scheres SHW. 2016. Resolving macromolecular structures from electron cryotomography data using subtomogram averaging in RELION. Nat. Protoc 11(11):2054–65 [PubMed: 27685097]
15. Bieber A, Capitanio C, Erdmann PS, Fiedler F, Beck F, et al. 2022. In situ structural analysis reveals membrane shape transitions during autophagosome formation. PNAS 119(39):e2209823119 [PubMed: 36122245]
16. Bouvette J, Liu H-F, Du X, Zhou Y, Sikkema AP, et al. 2021. Beam image-shift accelerated data acquisition for near-atomic resolution single-particle cryo-electron tomography. Nat. Commun 12:1957 [PubMed: 33785757]
17. Briggs JAG. 2013. Structural biology in situ—the potential of subtomogram averaging. Curr. Opin. Struct. Biol 23(2):261–67 [PubMed: 23466038]
18. Buchholz T-O, Jordan M, Pigino G, Jug F. 2019. Cryo-CARE: content-aware image restoration for cryo-transmission electron microscopy data. arXiv:1810.05420 [cs.CV]
19. Buchholz T-O, Krull A, Shahidi R, Pigino G, Jékely G, Jug F. 2019. Content-aware image restoration for electron microscopy. Methods Cell Biol 152:277–89 [PubMed: 31326025]
20. Burt A, Cassidy CK, Ames P, Bacia-Verloop M, Baulard M, et al. 2020. Complete structure of the chemosensory array core signalling unit in an *E. coli* minicell strain. Nat. Commun 11:743 [PubMed: 32029744]
21. Burt A, Gaifas L, Dendooven T, Gutsche I. 2021. A flexible framework for multi-particle refinement in cryo-electron tomography. PLOS Biol 19(8):e3001319 [PubMed: 34437530]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

22. Bykov YS, Schaffer M, Dodonova SO, Albert S, Plitzko JM, et al. 2017. The structure of the COPI coat determined within the cell. eLife 6:e32493 [PubMed: 29148969]
23. Cai S, Böck D, Pilhofer M, Gan L. 2018. The in situ structures of mono-, di-, and trinucleosomes in human heterochromatin. Mol. Biol. Cell 29(20):2450–57 [PubMed: 30091658]
24. Castaño-Díez D, Kudryashev M, Arheit M, Stahlberg H. 2012. Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. J. Struct. Biol 178(2):139–51 [PubMed: 22245546]
25. Chaikeeratisak V, Nguyen K, Khanna K, Brilot AF, Erb ML, et al. 2017. Assembly of a nucleus-like structure during viral replication in bacteria. Science 355(6321):194–97 [PubMed: 28082593]
26. Chen M, Dai W, Sun SY, Jonasch D, He CY, et al. 2017. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nat. Methods 14(10):983–85 [PubMed: 28846087]
27. Cheng A, Eng ET, Alink L, Rice WJ, Jordan KD, et al. 2018. High resolution single particle cryoelectron microscopy using beam-image shift. J. Struct. Biol 204(2):270–75 [PubMed: 30055234]
28. Chreifi G, Chen S, Metskas LA, Kaplan M, Jensen GJ. 2019. Rapid tilt-series acquisition for electron cryotomography. J. Struct. Biol 205(2):163–69 [PubMed: 30639925]
29. Clarke NI, Royle SJ. 2018. FerriTag is a new genetically-encoded inducible tag for correlative light-electron microscopy. Nat. Commun 9:2604 [PubMed: 29973588]
30. Collado J, Kalemanov M, Campelo F, Bourgoint C, Thomas F, et al. 2019. Tricalbin-mediated contact sites control ER curvature to maintain plasma membrane integrity. Dev. Cell 51(4):476–87.e7 [PubMed: 31743662]
31. Croxford M, Elbaum M, Arigovindan M, Kam Z, Agard D, et al. 2021. Entropy-regularized deconvolution of cellular cryotransmission electron tomograms. PNAS 118(50):e2108738118 [PubMed: 34876518]
32. Dahan I, Sorrentino S, Boujemaa-Paterski R, Medalia O. 2018. Tiopronin-protected gold nanoparticles as a potential marker for cryo-EM and tomography. Structure 26(10):1408–13.e3 [PubMed: 30078643]
33. Dahlberg PD, Perez D, Hecksel CW, Chiu W, Moerner WE. 2022. Metallic support films reduce optical heating in cryogenic correlative light and electron tomography. J. Struct. Biol 214(4):107901 [PubMed: 36191745]
34. Dahlberg PD, Saurabh S, Sartor AM, Wang J, Mitchell PG, et al. .2020. Cryogenic single-molecule fluorescence annotations for electron tomography reveal in situ organization of key proteins in Caulobacter. PNAS 117(25):13937–44 [PubMed: 32513734]
35. Danev R, Baumeister W. 2016. Cryo-EM single particle analysis with the Volta phase plate. eLife 5:e13046 [PubMed: 26949259]
36. DeRosier DJ. 2021. Where in the cell is my protein? Q. Rev. Biophys 54:e9 [PubMed: 34165063]
37. Dimchev G, Amiri B, Fäßler F, Falcke M, Schur FK. 2021. Computational toolbox for ultrastructural quantitative analysis of filament networks in cryo-ET data. J. Struct. Biol 213(4):107808 [PubMed: 34742832]
38. Dodonova SO, Aderhold P, Kopp J, Ganeva I, Röhling S, et al. 2017. 9Å structure of the COPI coat reveals that the Arf1 GTPase occupies two contrasting molecular environments. eLife 6:e26691 [PubMed: 28621666]
39. Douglas JO, Conroy M, Giuliani F, Gault B. 2022. In-situ sputtering from the micromanipulator to enable cryogenic preparation of specimens for atom probe tomography by focused-ion beam. arXiv:2211.06877 [cond-mat.mtrl-sci]
40. Dubochet J, Adrian M, Chang JJ, Homo JC, Lepault J, et al. 1988.Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys 21(2):129–228 [PubMed: 3043536]
41. Eisenstein F, Danev R, Pilhofer M. 2019. Improved applicability and robustness of fast cryo-electron tomography data acquisition. J. Struct. Biol 208(2):107–14 [PubMed: 31425790]
42. Eisenstein F, Yanagisawa H, Kashihara H, Kikkawa M, Tsukita S, Danev R. 2022. Parallel cryo electron tomography on in situ lamellae.bioRxiv 2022.04.07.487557.10.1101/2022.04.07.487557
43. Elbaum M, Seifer S, Houben L, Wolf SG, Rez P. 2021. Toward compositional contrast by cryo-STEM. Acc. Chem. Res 54(19):3621–31 [PubMed: 34491730]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

44. Elferich J, Schiroli G, Scadden D, Grigorieff N. 2022. Defocus corrected large area cryo-EM (DeCo-LACE) for label-free detection of molecules across entire cell sections. eLife 11:e80980 [PubMed: 36382886]
45. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74 [PubMed: 22955616]
46. Engel BD, Schaffer M, Albert S, Asano S, Plitzko JM, Baumeister W. 2015. In situ structural analysis of Golgi intracisternal protein arrays. PNAS 112(36):11264–69 [PubMed: 26311849]
47. Erdmann PS, Hou Z, Klumpe S, Khavnekar S, Beck F, et al. 2021. In situ cryo-electron tomography reveals gradient organization of ribosome biogenesis in intact nucleoli. Nat. Commun 12:5364 [PubMed: 34508074]
48. Fäßler F, Dimchev G, Hodirnau V-V, Wan W, Schur FKM. 2020. Cryo-electron tomography structure of Arp2/3 complex in cells reveals new insights into the branch junction. Nat. Commun 11:6437 [PubMed: 33353942]
49. Fäßler F, Zens B, Hauschild R, Schur FKM. 2020. 3D printed cell culture grid holders for improved cellular specimen preparation in cryo-electron microscopy. J. Struct. Biol 212(3):107633 [PubMed: 32987119]
50. Fatmaoui F, Carrivain P, Grewe D, Jacob B, Victor J-M, et al. 2022. Cryo-electron tomography and deep learning-based denoising reveal native chromatin landscapes of interphase nuclei. bioRxiv 2022.08.16.502515. 10.1101/2022.08.16.502515
51. Fernández JJ,Li S,Crowther RA.2006.CTF determination and correction in electron cryotomography. Ultramicroscopy 106(7):587–96 [PubMed: 16616422]
52. Ferreira JL, Pražák V, Vasishtan D, Siggel M, Hentzschel F, et al. 2022. Form follows function: variable microtubule architecture in the malaria parasite. bioRxiv 2022.04.13.488170. 10.1101/2022.04.13.488170
53. Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, et al. 2022. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. Science 376(6598):eabm9326 [PubMed: 35679401]
54. Foster HE, Ventura Santos C, Carter AP. 2022. A cryo-ET survey of microtubules and intracellular compartments in mammalian axons. J. Cell Biol 221(2):e202103154 [PubMed: 34878519]
55. Frangakis AS. 2021. It's noisy out there! A review of denoising techniques in cryo-electron tomography. J. Struct. Biol 213(4):107804 [PubMed: 34732363]
56. Frangakis AS,Hegerl R.2001.Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. J. Struct. Biol 135(3):239–50 [PubMed: 11722164]
57. Frank J. 2006.Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford, UK: Oxford Univ. Press
58. Frank J 2008. Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell. Berlin: Springer
59. Freitas RA Jr., Merkle RC. 2004. Kinematic Self-Replicating Machines. Georgetown, TX: Landes Biosci.
60. Friedman JR, Lackner LL, West M, DiBenedetto JR, Nunnari J, Voeltz GK. 2011. ER tubules mark sites of mitochondrial division. Science 334(6054):358–62 [PubMed: 21885730]
61. Giannuzzi LA. 2006. Introduction to Focused Ion Beams: Instrumentation, Theory, Techniques and Practice. Berlin: Springer
62. Glaeser RM. 2013. Invited review article: methods for imaging weak-phase objects in electron microscopy. Rev. Sci. Instrum 84(11):111101 [PubMed: 24289381]
63. Gorelick S, Buckley G, Gervinskas G, Johnson TK, Handley A, et al. 2019. PIE-scope, integrated cryo-correlative light and FIB/SEM microscopy. eLife 8:e45919 [PubMed: 31259689]
64. Grange M, Vasishtan D, Grünewald K. 2017. Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes. J. Struct. Biol 197(2):181–90 [PubMed: 27374320]
65. Grant T, Grigorieff N. 2015. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. eLife 4:e06980 [PubMed: 26023829]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

66. Grimm R, Typke D, Bärmann M, Baumeister W. 1996. Determination of the inelastic mean free path in ice by examination of tilted vesicles and automated most probable loss imaging. Ultramicroscopy 63(3–4):169–79 [PubMed: 8921626]

67. Hagen WJH, Wan W, Briggs JAG. 2017. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. J. Struct. Biol 197(2):191–98 [PubMed: 27313000]

68. Han D, Pal S, Nangreave J, Deng Z, Liu Y, Yan H. 2011. DNA origami with complex curvatures in three-dimensional space. Science 332(6027):342–46 [PubMed: 21493857]

69. Himes BA, Zhang P. 2018. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nat. Methods 15(11):955–61 [PubMed: 30349041]

70. Hoffman DP, Shtengel G, Xu CS, Campbell KR, Freeman M, et al. 2020. Correlative three-dimensional super-resolution and block-face electron microscopy of whole vitreously frozen cells.Science 367(6475):eaaz5357 [PubMed: 31949053]

71. Hoffmann PC, Bharat TAM, Wozny MR, Boulanger J, Miller EA, Kukulski W. 2019. Tricalbins contribute to cellular lipid flux and form curved ER-PM contacts that are bridged by rod-shaped structures. Dev. Cell 51(4):488–502.e8 [PubMed: 31743663]

72. Iancu CV, Tivol WF, Schooler JB, Dias DP, Henderson GP, et al. 2006. Electron cryotomography sample preparation using the Vitrobot. Nat. Protoc 1(6):2813–19 [PubMed: 17406539]

73. Jasnin M, Ecke M, Baumeister W, Gerisch G. 2016. Actinorganization in cells responding to a perforated surface, revealed by live imaging and cryo-electron tomography. Structure 24(7):1031–43 [PubMed: 27320835]

74. Jiang W, Wagner J, Du W, Plitzko J, Baumeister W, et al. 2022. A transformation clustering algorithm and its application in polyribosomes structural profiling. Nucleic Acids Res 50(16):9001–11 [PubMed: 35811088]

75. Kelley K, Raczkowski AM, Klykov O, Jaroenlak P, Bobe D, et al. 2022. Waffle method: a general and flexible approach for improving throughput in FIB-milling. Nat. Commun 13:1857 [PubMed: 35387991]

76. Khanna K, Villa E. 2022. Revealing bacterial cell biology using cryo-electron tomography. Curr. Opin. Struct. Biol 75:102419 [PubMed: 35820259]

77. Khavnekar S, Wan W, Majumder P, Wietrzynski W, Erdmann PS, Plitzko JM. 2023. Multishot tomography for high-resolution in situ subtomogram averaging. J. Struct. Biol 215:107911 [PubMed: 36343843]

78. Khavnekar S, Vrbovská V, Zaoralová M, Kelley R, Beck F, et al. 2022. Optimizing cryo-FIB lamellas for sub-5Å in situ structural biology. bioRxiv 2022.06.16.496417. 10.1101/2022.06.16.496417

79. Klumpe S, Fung HK, Goetz SK, Zagoriy I, Hampoelz B, et al. 2021. A modular platform for automated cryo-FIB workflows. eLife 10:e70506 [PubMed: 34951584]

80. Kuba J, Mitchels J, Hovorka M, Erdmann P, Berka L, et al. 2021. Advanced cryo-tomography workflow developments—correlative microscopy, milling automation and cryo-lift-out. J. Microsc 281(2):112–24 [PubMed: 32557536]

81. Kühlbrandt W 2014. The resolution revolution. Science 343(6178):1443–44 [PubMed: 24675944]

82. Lam V, Villa E. 2021. Practical approaches for cryo-FIB milling and applications for cellular cryo-electron tomography. Methods Mol. Biol 2215:49–82 [PubMed: 33367999]

83. Lamm L, Righetto RD, Wietrzynski W, Pöge M, Martinez-Sanchez A, et al. 2022. MemBrain: a deep learning-aided pipeline for automated detection of membrane proteins in cryo-electron tomograms. Comput. Methods Programs Biomed 224:106990 [PubMed: 35858496]

84. Laughlin TG, Deep A, Prichard AM, Seitz C, Gu Y, et al. 2022. Architecture and self-assembly of the jumbo bacteriophage nuclear shell. Nature 608(7922):429–35 [PubMed: 35922510]

85. Lazić I, Wirix M, Leidl ML, de Haas F, Mann D, et al. 2022. Single-particle cryo-EM structures from iDPC-STEM at near-atomic resolution. Nat. Methods 19(9):1126–36 [PubMed: 36064775]

86. Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, et al. 2018. Noise2Noise: learning image restoration without clean data. arXiv:1803.04189 [cs.CV]

87. Liu Y-T, Zhang H, Wang H, Tao C-L, Bi G-Q, Zhou ZH. 2021. Isotropic reconstruction of electron tomograms with deep learning. bioRxiv 2021.07.17.452128. 10.1101/2021.07.17.452128

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

88. Liu Y-T, Zhang H, Wang H, Tao C-L, Bi G-Q, Zhou ZH. 2022. Isotropic reconstruction for electron tomography with deep learning. Nat. Commun 13:6482 [PubMed: 36309499]

89. Lucas BA, Himes BA, Xue L, Grant T, Mahamid J, Grigorieff N. 2021. Locating macromolecular assemblies in cells by 2D template matching with cisTEM. eLife 10:e68946 [PubMed: 34114559]

90. Lucas BA, Zhang K, Loerch S, Grigorieff N. 2022. In situ single particle classification reveals distinct 60S maturation intermediates in cells. eLife 11:e79272 [PubMed: 36005291]

91. Mageswaran SK, Grotjahn DA, Zeng X, Barad BA, Medina M, et al. 2021. Nanoscale details of mitochondrial fission revealed by cryo-electron tomography. bioRxiv 2021.12.13.472487. 10.1101/2021.12.13.472487

92. Mahamid J, Pfeffer S, Schaffer M, Villa E, Danev R, et al. 2016. Visualizing the molecular sociology at the HeLa cell nuclear periphery. Science 351(6276):969–72 [PubMed: 26917770]

93. Mahamid J,Schampers R,Persoon H,Hyman AA,Baumeister W,Plitzko JM.2015.A focused ion beam milling and lift-out approach for site-specific preparation of frozen-hydrated lamellas from multicellular organisms. J. Struct. Biol 192(2):262–69 [PubMed: 26216184]

94. Marko M, Hsieh C, Schalek R, Frank J, Mannella C. 2007. Focused-ion-beam thinning of frozen-hydrated biological specimens for cryo-electron microscopy. Nat. Methods 4(3):215–17 [PubMed: 17277781]

95. Martinez-Sanchez A, Garcia I, Asano S, Lucic V, Fernandez J-J. 2014. Robust membrane detection based on tensor voting for electron tomography. J. Struct. Biol 186(1):49–61 [PubMed: 24625523]

96. Martinez-Sanchez A, Kochovski Z, Laugks U, Meyer Zum Alten Borgloh J, Chakraborty S, et al. 2020. Template-free detection and classification of membrane-bound complexes in cryo-electron tomograms. Nat. Methods 17(2):209–16 [PubMed: 31907446]

97. Martynowycz MW, Clabbers MTB, Unge J, Hattne J, Gonen T. 2021. Benchmarking the ideal sample thickness in cryo-EM. PNAS 118(49):e210884118

98. Mastronarde DN. 2005. Automated electron microscope tomography using robust prediction of specimen movements. J. Struct. Biol 152(1):36–51 [PubMed: 16182563]

99. Mastronarde DN. 2013. Automated tomographic reconstruction in the IMOD software package. Microsc. Microanal 19(S2):544–45 [PubMed: 23534938]

100. Mastronarde DN, Held SR. 2017. Automated tilt series alignment and tomographic reconstruction in IMOD. J. Struct. Biol 197(2):102–13 [PubMed: 27444392]

101. Mattei S, Glass B, Hagen WJH, Kräusslich H-G, Briggs JAG. 2016. The structure and flexibility of conical HIV-1 capsids determined within intact virions. Science 354(6318):1434–37 [PubMed: 27980210]

102. Milo R, Jorgensen P, Moran U, Weber G, Springer M. 2010. BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res 38:D750–53 [PubMed: 19854939]

103. Milo R, Philips R. 2015. Cell Biology by the Numbers. New York: Garland Sci.

104. Moebel E, Martinez-Sanchez A, Lamm L, Righetto RD, Wietrzynski W, et al. 2021. Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms. Nat. Methods 18(11):1386–94 [PubMed: 34675434]

105. Mosalaganti S, Kosinski J, Albert S, Schaffer M, Strenkert D, et al. 2018. In situ architecture of the algal nuclear pore complex. Nat. Commun 9:2361 [PubMed: 29915221]

106. Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, et al. 2022. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. Science 376(6598):eabm9506 [PubMed: 35679397]

107. Ni T, Frosio T, Mendonça L, Sheng Y, Clare D, et al. 2022. High-resolution in situ structure determination by cryo-electron tomography and subtomogram averaging using emClarity. Nat. Protoc 17(2):421–44 [PubMed: 35022621]

108. O'Reilly FJ, Xue L, Graziadei A, Sinn L, Lenz S, et al. 2020. In-cell architecture of an actively transcribing-translating expressome. Science 369(6503):554–57 [PubMed: 32732422]

109. Ori A, Banterle N, Iskar M, Andrés-Pons A, Escher C, et al. 2013. Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. Mol. Syst. Biol 9:648 [PubMed: 23511206]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

110. Peck A, Carter SD, Mai H, Chen S, Burt A, Jensen GJ. 2022. Montage electron tomography of vitrified specimens. J. Struct. Biol 214:107860 [PubMed: 35487464]
111. Peck JV, Fay JF, Strauss JD. 2022. High-speed high-resolution data collection on a 200 keV cryo-TEM. IUCrJ 9(Pt 2):243–52
112. Phan S, Boassa D, Nguyen P, Wan X, Lanman J, et al. 2017. 3D reconstruction of biological structures: automated procedures for alignment and reconstruction of multiple tilt series in electron tomography. Adv. Struct. Chem. Imaging 2:8 [PubMed: 27547706]
113. Phan S, Terada M, Lawrence A. 2009. Serial reconstruction and montaging from large-field electron microscope tomograms. Conf. Proc. IEEE Eng. Med. Biol. Sci 2009:5772–76
114. Piche N, Bouchard I, Marsh M. 2017. Dragonfly Segmentation Trainer—a general and user-friendly machine learning image segmentation solution. Microsc. Microanal 23(S1):132–33
115. Prinz WA, Toulmay A, Balla T. 2020. The functional universe of membrane contact sites. Nat. Rev. Mol. Cell Biol 21(1):7–24 [PubMed: 31732717]
116. Pyle E, Hutchings J, Zanetti G. 2022. Strategies for picking membrane-associated particles within subtomogram averaging workflows. Faraday Discuss 240:101–13 [PubMed: 35924570]
117. Pyle E, Zanetti G. 2021. Current data processing strategies for cryo-electron tomography and subtomogram averaging. Biochem. J 478(10):1827–45 [PubMed: 34003255]
118. Radermacher M 1988. Three-dimensional reconstruction of single particles from random and nonrandom tilt series. J. Electron Microsc. Technol 9(4):359–94
119. Rice WJ, Cheng A, Noble AJ, Eng ET, Kim LY, et al. 2018. Routine determination of ice thickness for cryo-EM grids. J. Struct. Biol 204(1):38–44 [PubMed: 29981485]
120. Rickgauer JP, Grigorieff N, Denk W. 2017. Single-protein detection in crowded molecular environments in cryo-EM images. eLife 6:e25648 [PubMed: 28467302]
121. Sanchez RM, Zhang Y, Chen W, Dietrich L, Kudryashev M. 2020. Subnanometer-resolution structure determination in situ by hybrid subtomogram averaging—single particle cryo-EM. Nat. Commun 11:3709 [PubMed: 32709843]
122. Sarder P, Nehorai A. 2006. Deconvolution methods for 3-D fluorescence microscopy images. IEEE Signal Process. Mag 23(3):32–45
123. Schaffer M, Mahamid J, Engel BD, Laugks T, Baumeister W, Plitzko JM. 2017. Optimized cryo-focused ion beam sample preparation aimed at in situ structural studies of membrane proteins. J. Struct. Biol 197(2):73–82 [PubMed: 27444390]
124. Schmid MF, Booth CR. 2008. Methods for aligning and for averaging 3D volumes with missing data. J. Struct. Biol 161(3):243–48 [PubMed: 18299206]
125. Schorb M, Briggs JAG. 2014. Correlated cryo-fluorescence and cryo-electron microscopy with high spatial precision and improved sensitivity. Ultramicroscopy 143:24–32 [PubMed: 24275379]
126. Schuller AP, Wojtynek M, Mankus D, Tatli M, Kronenberg-Tenga R, et al. 2021. The cellular environment shapes the nuclear pore complex architecture. Nature 598:667–71 [PubMed: 34646014]
127. Schur FKM, Obr M, Hagen WJH, Wan W, Jakobi AJ, et al. .2016.An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. Science 353(6298):506–8 [PubMed: 27417497]
128. Schwartz O, Axelrod JJ, Campbell SL, Turnbaugh C, Glaeser RM, Müller H. 2019. Laser phase plate for transmission electron microscopy. Nat. Methods 16(10):1016–20 [PubMed: 31562475]
129. Sergey G, Denis K, Ava H, Gediminas G, Viola O, et al. 2018. Oxygen plasma focused ion beam scanning electron microscopy for biological samples. bioRxiv 457820. 10.1101/457820
130. Serwas D, Akamatsu M, Moayed A, Vegesna K, Vasan R, et al. 2021. Actin force generation in vesicle formation: mechanistic insights from cryo-electron tomography. Dev. Cell 57:P1132–45.E5
131. Song K, Shang Z, Fu X, Lou X, Grigorieff N, Nicastro D. 2020. In situ structure determination at nanometer resolution using TYGRESS. Nat. Methods 17(2):201–8 [PubMed: 31768058]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

132. Stalling D, Westerhoff M, Hege H-C. 2005. Amira: a highly interactive system for visual data analysis. In The Visualization Handbook, ed. Hansen C, Johnson C, pp. 749–67. Amsterdam: Elsevier
133. Studer D, Humbel BM, Chiquet M. 2008. Electron microscopy of high pressure frozen samples: bridging the gap between cellular ultrastructure and atomic resolution. Histochem. Cell Biol 130(5):877–89 [PubMed: 18795316]
134. Suloway C, Shi J, Cheng A, Pulokas J, Carragher B, et al. 2009. Fully automated, sequential tilt-series acquisition with Leginon. J. Struct. Biol 167(1):11–18 [PubMed: 19361558]
135. Sun SY, Segev-Zarko L-A, Chen M, Pintilie GD, Schmid MF, et al. 2022. Cryo-ET of *Toxoplasma* parasites gives subnanometer insight into tubulin-based structures. PNAS 119(6):e2111661119 [PubMed: 35121661]
136. Tacke S, Erdmann P, Wang Z, Klumpe S, Grange M, et al. 2021. A streamlined workflow for automated cryo focused ion beam milling. J. Struct. Biol 213(3):107743 [PubMed: 33971286]
137. Tegunov D, Xue L, Dienemann C, Cramer P, Mahamid J. 2021. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. Nat. Methods 18(2):186–93 [PubMed: 33542511]
138. Toro-Nahuelpan M, Zagoriy I, Senger F, Blanchoin L, Théry M, Mahamid J. 2020. Tailoring cryo-electron microscopy grids by photo-micropatterning for in-cell structural studies. Nat. Methods 17(1):50–54 [PubMed: 31740821]
139. Turoňová B, Hagen WJH, Obr M, Mosalaganti S, Beugelink JW, et al. 2020. Benchmarking tomographic acquisition schemes for high-resolution structural biology. Nat. Commun 11:876 [PubMed: 32054835]
140. Turoňová B, Schur FKM, Wan W, Briggs JAG. 2017. Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4Å. J. Struct. Biol 199(3):187–95 [PubMed: 28743638]
141. Villa E, Schaffer M, Plitzko JM, Baumeister W. 2013.Opening windows into the cell: focused-ion-beam milling for cryo-electron tomography. Curr. Opin. Struct. Biol 23(5):771–77 [PubMed: 24090931]
142. Verma RS. 1995. Genes V. by Benjamin Lewin Oxford University Press, Oxford and New York, 1994, 1,272 pp. Am. J. Med. Genet 57:121
143. Wagner FR, Watanabe R, Schampers R, Singh D, Persoon H, et al. 2020. Preparing samples from whole cells using focused-ion-beam milling for cryo-electron tomography. Nat. Protoc 15(6):2041–70 [PubMed: 32405053]
144. Wagner J, Schaffer M, Fernández-Busnadiego R. 2017. Cryo-electron tomography—the cell biology that came in from the cold. FEBS Lett 591(17):2520–33 [PubMed: 28726246]
145. Wang K, Strunk K, Zhao G, Gray JL, Zhang P. 2012. 3D structure determination of native mammalian cells using cryo-FIB and cryo-electron tomography. J. Struct. Biol 180(2):318–26 [PubMed: 22796867]
146. Wang Q, Mercogliano CP, Löwe J. 2011. A ferritin-based label for cellular electron cryotomography. Structure 19(2):147–54 [PubMed: 21300284]
147. Wan W, Briggs JAG. 2016. Cryo-electron tomography and subtomogram averaging. Methods Enzymol 579:329–67 [PubMed: 27572733]
148. Wan W, Khavnekar S, Wagner J, Erdmann P, Baumeister W. 2020. STOPGAP: a software package for subtomogram averaging and refinement. Microsc. Microanal 26(S2):2516
149. Watanabe R, Buschauer R, Böhning J, Audagnotto M, Lasker K, et al. 2020. The in situ structure of Parkinson's disease-linked LRRK2. Cell 182:1508–18.e16 [PubMed: 32783917]
150. Waugh B, Wolf SG, Fass D, Branlund E, Kam Z, et al. 2020. Three-dimensional deconvolution processing for STEM cryotomography. PNAS 117(44):27374–80 [PubMed: 33077585]
151. Weis F, Hagen WJH. 2020. Combining high throughput and high quality for cryo-electron microscopy data collection. Acta Crystallogr. D 76(Pt 8):724–28
152. Weiss GL, Kieninger A-K, Maldener I, Forchhammer K, Pilhofer M. 2019. Structure and function of a bacterial gap junction analog. Cell 178(2):374–84.e15 [PubMed: 31299201]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

153. Wietrzynski W, Schaffer M, Tegunov D, Albert S, Kanazawa A, et al. 2020. Charting the native architecture of *Chlamydomonas* thylakoid membranes with single-molecule precision. eLife 9:e53740 [PubMed: 32297859]

154. Winey M, Meehl JB, O'Toole ET, Giddings TH Jr. 2014. Conventional transmission electron microscopy. Mol. Biol. Cell 25(3):319–23 [PubMed: 24482357]

155. Winey M, Yarar D, Giddings TH Jr., Mastronarde DN. 1997. Nuclear pore complex number and distribution throughout the *Saccharomyces cerevisiae* cell cycle by three-dimensional reconstruction from electron micrographs of nuclear envelopes. Mol. Biol. Cell 8:2119–32 [PubMed: 9362057]

156. Wolf SG, Houben L, Elbaum M. 2014. Cryo-scanning transmission electron tomography of vitrified cells. Nat. Methods 11(4):423–28 [PubMed: 24531421]

157. Wolff G, Limpens RWAL, Zheng S, Snijder EJ, Agard DA, et al. 2019. Mind the gap: Micro-expansion joints drastically decrease the bending of FIB-milled cryo-lamellae. J. Struct. Biol 208(3):107389 [PubMed: 31536774]

158. Wozny MR, Di Luca A, Morado DR, Picco A, Hoffmann PC, et al. 2022. Supramolecular architecture of the ER-mitochondria encounter structure in its native environment. bioRxiv 2022.04.12.488000. 10.1101/2022.04.12.488000

159. Xing J, Takeuchi K, Kamei K, Nakamuro T, Harano K, Nakamura E. 2022. Atomic-number (Z)-correlated atomic sizes for deciphering electron microscopic molecular images. PNAS 119(14):e2114432119 [PubMed: 35349339]

160. Yamada H, Yamaguchi M, Shimizu K, Murayama SY, Mitarai S, et al. 2017. Structome analysis of *Escherichia coli* cells by serial ultrathin sectioning reveals the precise cell profiles and the ribosome density. Microscopy 66:283–94 [PubMed: 28854579]

161. Yang JE, Larson MR, Sibert BS, Kim JY, Parrell D, et al. 2022. Correlative cryogenic montage electron tomography for comprehensive in-situ whole-cell structural studies. bioRxiv 2021.12.31.474669. 10.1101/2021.12.31.474669

162. Yesibolati MN, Laganá S, Kadkhodazadeh S, Mikkelsen EK, Sun H, et al. 2020. Electron inelastic mean free path in water. Nanoscale 12(40):20649–57 [PubMed: 32614016]

163. Zachs T, Schertel A, Medeiros J, Weiss GL, Hugener J, et al. 2020. Fully automated, sequential focused ion beam milling for cryo-electron tomography. eLife 9:e52286 [PubMed: 32149604]

164. Zhang P 2019. Advances in cryo-electron tomography and subtomogram averaging and classification. Curr. Opin. Struct. Biol 58:249–58 [PubMed: 31280905]

165. Zheng S, Wolff G, Greenan G, Chen Z, Faas FGA, et al. 2022. Are Tomo: an integrated software package for automated marker-free, motion-corrected cryo-electron tomographic alignment and reconstruction. J. Struct. Biol. X 6:100068 [PubMed: 35601683]

166. Zimmerli CE, Allegretti M, Rantos V, Goetz SK, Obarska-Kosinska A, et al. 2021. Nuclear pores dilate and constrict in cellulo. Science 374(6573):eabd9776 [PubMed: 34762489]

167. Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJ, et al. 2018. New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife 7:e42166 [PubMed: 30412051]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

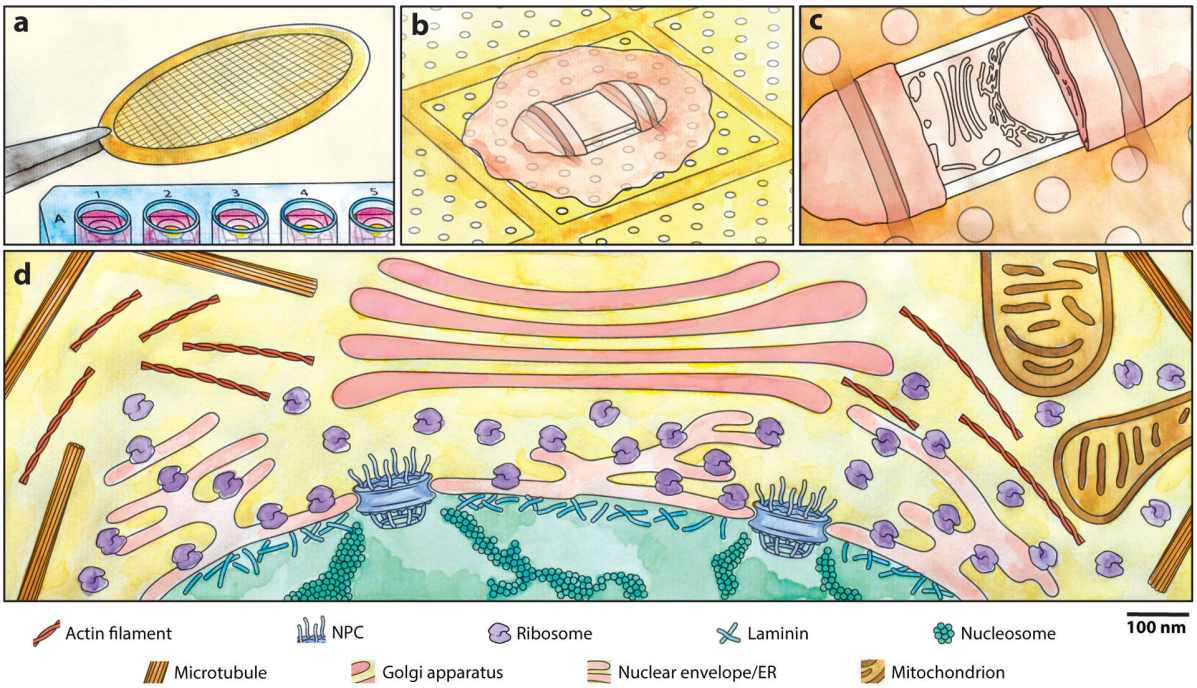


Figure 1.

Cryo-FIB-ET workflow. (*a*) Adherent cells can be cultured on EM grids, allowing unique lab-on-a-grid experiments. (*b*) A thin window or lamella is carved into an adherent grid grown on an EM grid via an FIB under cryogenic conditions. Expansion joints (trenches) on either side of the lamella are made to improve mechanical stability by offering points to allow flexing. (*c*) The lamella is imaged by TEM under cryogenic conditions at multiple tilt angles to facilitate a 3D reconstruction of the molecules within the lamella. (*d*) The complex and heterogeneous interior of a cell can be imaged at the highest achievable spatial resolution, resolving individual macromolecules (ribosomes, the NPC, nucleosomes), filaments (microtubules, actin, lamin), and organelles (mitochondria, ER, nucleus, Golgi). Abbreviations: cryo-FIB-ET, cryo-focused ion beam milling and electron tomography; EM, electron microscopy; ER, endoplasmic reticulum; FIB, focused ion beam; NPC, nuclear pore complex; TEM, transmission electron microscopy. Figure copyright 2023 Dorotea Fracchiolla, adapted with permission.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

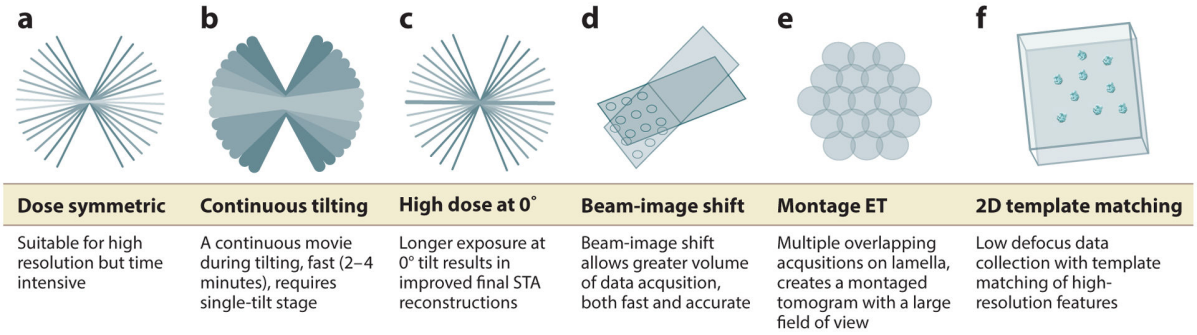


Figure 2.

Data collection methods. What data collection method is best for your data? It depends. (*a*) In the dose-symmetric tilt series, data collection alternates between positive and negative tilts. (*b*) In continuous fast tilting, data are continuously acquired. (*c*) In hybrid STA-SPA, a high-dose image is collected at 0°, followed by implementation of the dose-symmetric method. (*d*) In beam-image shift, the beam is shifted from a central focus and tracking to targeted areas of acquisition, facilitating the collection of multiple tilt series simultaneously. (*e*) In a montage collection series, the entire lamella is imaged to create one large tomogram. (*f*) In 2D template matching, one single projection image is acquired at 0°, and high-resolution features of a target are matched to the data. Abbreviations: ET, electron tomography; SPA, single-particle analysis; STA, subtomogram analysis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Macromolecule	Abundance	Size	Sample datasets for solved structures			
			Lamella(e)	Tomograms	Particles	Resolution
Nuclear pore complex						
Mammalian	10^4	125 MDa	153 lamellae	54 tomograms	194 particles	~34 Å
S. pombe WT condition	10^2	66 MDa	~3.9 NPCs per tomograms	78 tomograms	2,237 particles	22 Å
Ribosome						
Mammalian (80S)	6×10^6	4.5 MDa	17 lamellae	180 tomograms	15,628 particles	4.9 Å
Bacterial (70S)	2.6×10^4	2.5 MDa	---	65 tomograms	17,890 particles	3.5 Å
Nucleosomes						
Mammalian	3×10^7/nucleus	200 kDa	1 lamella	1 tomogram	1,141 particles	24 Å
COPII		600 kDa	60 lamellae	60 tomograms	1,866 particles	9 Å
Arp2/3		224 kDa	---	131 tomograms	14,296 particles	9 Å

Figure 3.

Cryo-ET by the numbers. The size and cellular abundance of a given macromolecule (or its local concentration) are relevant when estimating how many lamellae, tomograms, and particles you will need to perform STA. Not surprisingly, the more abundant protein complexes (e.g., ribosomes, nucleosomes) will have a greater ratio of particles to tomograms compared to less abundant complexes (e.g., the NPC). For the number of NPCs in yeast, see Reference 155; particle numbers and estimated resolution from EMDB 11373 (166). For the number of NPCs in mammals, see Reference 109; particle numbers and estimated resolution from EMDB 12814 (126). For the number of nucleosomes in the mammalian nucleus, see Reference 45; particle numbers and estimated resolution from EMDB 6949 (23). For the number of bacterial ribosomes, see References 59 and 142; BioID 106861; particle numbers and estimated resolution from EMDB 11650 (137). For the number of mammalian ribosomes, see Reference 160; BioID 106861; particle numbers and estimated resolution from Reference 78 and EMDB XXXX (yet to be released). COPII particle numbers and estimated resolution from EMDB 3720 (38). Arp2/3 particle numbers and estimated resolution from EMDB 11869 (48). Abbreviations: ET, electron tomography; NPC, nuclear pore complex; *S. pombe*, *Saccharomyces pombe*; STA, subtomogram analysis; WT, wild type.

E Wikipedia index

Wikipedia is often the most up-to-date overview of a topic, and much of the text of this book is adapted from it. For each main section, this index lists the most relevant articles with the opening sentences of each article (CC BY-SA 4.0). Wikipedia articles change; check the date and the references of an article before relying on it.

E.1 Sequence searching

Wikipedia articles for Chapter 1.

E.1.1 Homology and sequence similarity

- **Homology (biology)** — In biology, homology is similarity in anatomical structures or genes between organisms of different taxa due to shared ancestry, regardless of current functional differences.
- **Sequence homology** — Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life.
- **Gene** — In biology, the word gene has two meanings. The Mendelian gene is a basic unit of heredity. The molecular gene is a sequence of nucleotides in DNA that is transcribed to produce RNA.
- **Protein** — Proteins are large biomolecules and macromolecules that comprise one or more long chains of amino acid residues.
- **Convergent evolution** — Convergent evolution is the independent evolution of similar features in species of different lineages. Convergent evolution creates analogous structures that have similar form or function but were not present in the last common ancestor of those groups.
- **Gene duplication** — Gene duplication is a mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.
- **Horizontal gene transfer** — Horizontal gene transfer (HGT) or lateral gene transfer (LGT) is the movement of genetic material between organisms other than by the ("vertical") transmission of DNA from parent to offspring (reproduction).

E.1.2 Dynamic programming and sequence alignment

- **Dynamic programming** — Dynamic programming (DP) is both a mathematical optimization method and an algorithmic paradigm. The method was developed by Richard Bellman in the 1950s and has found applications in numerous fields, such as aerospace engineering and economics.
- **Sequence alignment** — In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

- **Needleman–Wunsch algorithm** — The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B.
- **Smith–Waterman algorithm** — The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences.
- **Gap penalty** — In DNA analysis, gap penalty is a method of scoring alignments of two or more DNA sequences. When aligning sequences, introducing gaps in the sequences can allow an alignment algorithm to match more terms than a gap-less alignment can.
- **Substitution matrix** — In bioinformatics and evolutionary biology, a substitution matrix describes the frequency at which a character in a nucleotide sequence or a protein sequence changes to other character states over evolutionary time.
- **BLOSUM** — In bioinformatics, the BLOSUM matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.
- **Point accepted mutation** — A point accepted mutation — also known as a PAM — is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection.

E.1.3 Heuristic search methods

- **BLAST (biotechnology)** — In bioinformatics, BLAST is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins, nucleotides of DNA and/or RNA sequences.
- **FASTA** — FASTA is a DNA and protein sequence alignment software package first described by David J. Lipman and William R. Pearson in 1985. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.
- **Gumbel distribution** — In probability theory and statistics, the Gumbel distribution is used to model the distribution of the maximum of a number of samples of various distributions.
- **Suffix array** — In computer science, a suffix array is a sorted array of all suffixes of a string. It is a data structure used in, among others, full-text indices, data-compression algorithms, and the field of bibliometrics.
- **Burrows–Wheeler transform** — The Burrows–Wheeler transform (BWT) rearranges a character string into runs of similar characters, in a manner that can be reversed to recover the original string.
- **FM-index** — In computer science, an FM-index is a compressed full-text substring index based on the Burrows–Wheeler transform, with some similarities to the suffix array.

E.1.4 Alignment-free sequence comparison

- **Alignment-free sequence analysis** — In bioinformatics, alignment-free sequence analysis approaches to molecular sequence and structure data provide alternatives over alignment-based approaches.
- **K-mer** — In bioinformatics, k-mers are substrings of length contained within a biological sequence.
- **MinHash** — In computer science and data mining, MinHash is a technique for quickly estimating how similar two sets are.
- **Jaccard index** — The Jaccard index is a statistic used for gauging the similarity and diversity of sample sets. It is defined in general taking the ratio of two sizes, the intersection size divided by the union size, also called intersection over union (IoU).

E.1.5 Profiles and multiple sequence alignments

- **Multiple sequence alignment** — Multiple sequence alignment (MSA) is the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.
- **Clustal** — Clustal is a computer program used for multiple sequence alignment in bioinformatics. It is one of the most widely cited bioinformatics software with two of its academic publications amongst the top 100 papers cited of all time, according to Nature in 2014.
- **Sequence motif** — In biology, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and usually assumed to be related to biological function of the macromolecule.
- **Position weight matrix** — A position weight matrix (PWM), also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), is a commonly used representation of motifs (patterns) in biological sequences.
- **Sequence logo** — In bioinformatics, a sequence logo is a graphical representation of the sequence conservation of nucleotides or amino acids . A sequence logo is created from a collection of aligned sequences and depicts the consensus sequence and diversity of the sequences.
- **Hidden Markov model** — In probability theory, a hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process. An HMM requires that there be an observable process whose outcomes depend on the outcomes of in a known way.
- **Viterbi algorithm** — The Viterbi algorithm is a dynamic programming algorithm that finds the most likely sequence of hidden events that would explain a sequence of observed events. The result of the algorithm is often called the Viterbi path.
- **HMMER** — HMMER is a free and commonly used software package for sequence analysis written by Sean Eddy. Its general usage is to identify homologous protein or nucleotide sequences, and to perform sequence alignments.

E.1.6 RNA homology search

- **Non-coding RNA** — A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein. The DNA sequence from which a functional non-coding RNA is transcribed is often called a non-coding gene or an RNA gene.
- **Nucleic acid secondary structure** — Nucleic acid secondary structure is the basepairing interactions within a single nucleic acid polymer or between two polymers. It can be represented as a list of bases which are paired in a nucleic acid molecule.
- **Nucleic acid structure prediction** — Nucleic acid structure prediction is a computational method to determine secondary and tertiary nucleic acid structure from its sequence. Secondary structure can be predicted from one or several nucleic acid sequences.
- **Probabilistic context-free grammar** — In theoretical linguistics and computational linguistics, probabilistic context free grammars (PCFGs) extend context-free grammars, similar to how hidden Markov models extend regular grammars. Each production is assigned a probability.
- **Rfam** — Rfam is a database containing information about non-coding RNA (ncRNA) families and other structured RNA elements.

E.1.7 Databases and domains

- **Protein domain** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure.
- **Protein family** — A protein family is a group of evolutionarily related proteins. In many cases, a protein family has a corresponding gene family, in which each gene encodes a corresponding protein with a 1:1 relationship.

- **Protein superfamily** — A protein superfamily is the largest grouping (clade) of proteins for which common ancestry can be inferred. Usually this common ancestry is inferred from structural alignment and mechanistic similarity, even if no sequence similarity is evident.
- **Pfam** — Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. The latest version of Pfam, 37.0, was released in June 2024 and contains 21,979 families.
- **InterPro** — InterPro is a database of protein families, protein domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterise them.
- **Structural Classification of Proteins database** — The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences.
- **CATH database** — The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.

E.1.8 Protein language models

- **Word embedding** — In natural language processing, a word embedding is a representation of a word. The embedding is used in text analysis.
- **Transformer (deep learning)** — In deep learning, the transformer is a family of artificial neural network architectures based on the multi-head attention mechanism, in which input data such as text, images, or audio is converted to a sequence of numerical representations called tokens, and each token is converted into a vector through an embedding layer.
- **BERT (language model)** — Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google. It learns to represent text as a sequence of vectors using self-supervised learning.
- **Large language model** — A large language model (LLM) is an AI model trained on a vast amount of text for natural language processing tasks, especially language generation. LLMs can typically generate, summarize, translate, and analyze text in many contexts.

E.1.9 Benchmarking search methods

- **Receiver operating characteristic** — A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier model at varying threshold values.
- **Precision and recall** — In classification, precision and recall are performance metrics that apply to data retrieved from a collection, corpus or sample space.
- **Sensitivity and specificity** — In medicine and statistics, sensitivity and specificity mathematically describe the accuracy of a test that reports the presence or absence of a medical condition.
- **False discovery rate** — In statistics, the false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons.
- **Structural Classification of Proteins database** — The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

E.1.10 Whole-genome alignment and conservation

- **Comparative genomics** — Comparative genomics is a branch of biological research that examines genome sequences across a spectrum of species, spanning from humans and mice to a diverse array of organisms from bacteria to chimpanzees.

- **Conserved sequence** — In evolutionary biology, conserved sequences are identical or similar sequences in nucleic acids or proteins across species, or within a genome, or between donor and receptor taxa.
- **Phylogenetic footprinting** — Phylogenetic footprinting is a technique used to identify transcription factor binding sites (TFBS) within a non-coding region of DNA of interest by comparing it to the orthologous sequence in different species.
- **Human accelerated regions** — Human accelerated regions (HARs), first described in August 2006, are a set of 49 segments of the human genome that are conserved throughout vertebrate evolution but are strikingly different in humans.

E.2 Databases

Wikipedia articles for Chapter 2.

E.2.1 How biological databases work

- **Biological database** — Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis.
- **Accession number (bioinformatics)** — An accession number, in bioinformatics, is a unique identifier given to a DNA or protein sequence record to allow for tracking of different versions of that sequence record and the associated sequence over time in a single data repository.
- **Entrez** — The Entrez Global Query Cross-Database Search System is a federated search engine, or web portal that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information (NCBI) website.
- **PubMed** — PubMed is an openly accessible, free database which primarily includes the MEDLINE database of references and abstracts on life sciences and biomedical topics.
- **Biocuration** — Biocuration is the field of life sciences dedicated to organizing biomedical data, information and knowledge into structured formats, such as spreadsheets, tables and knowledge graphs.
- **FAIR data** — FAIR data is data which meets the 2016 FAIR principles of findability, accessibility, interoperability, and reusability, first formally published in 2016.
- **PubMed Central** — PubMed Central (PMC) is a free digital repository that archives open access full-text scholarly articles that have been published in biomedical and life sciences journals.
- **Europe PubMed Central** — Europe PubMed Central is an open-access repository that contains millions of biomedical research works. It was known as UK PubMed Central until 1 November 2012.

E.2.2 Sequence databases

- **International Nucleotide Sequence Database Collaboration** — The International Nucleotide Sequence Database Collaboration (INSDC) consists of a joint effort to collect and disseminate databases containing DNA and RNA sequences.
- **GenBank** — The GenBank sequence database is an open access, annotated collection of publicly available nucleotide sequences and their protein translations.
- **European Nucleotide Archive** — The European Nucleotide Archive (ENA) is a repository providing free and unrestricted access to annotated DNA and RNA sequences.

- **Sequence Read Archive** — The Sequence Read Archive is a bioinformatics database that provides a public repository for DNA sequencing data, especially the "short reads" generated by high-throughput sequencing, which are typically less than 1,000 base pairs in length.
- **RefSeq** — The Reference Sequence (RefSeq) database is an open access, annotated and curated collection of publicly available nucleotide sequences and their protein products. RefSeq was introduced in 2000.
- **UniProt** — UniProt is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects.
- **Pfam** — Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. The latest version of Pfam, 37.0, was released in June 2024 and contains 21,979 families.
- **InterPro** — InterPro is a database of protein families, protein domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterise them.

E.2.3 Genome databases and browsers

- **Ensembl genome database project** — Ensembl genome database project is a scientific project at the European Bioinformatics Institute, which provides a centralized resource for geneticists, molecular biologists and other researchers studying the genomes of our own species and other vertebrates and model organisms.
- **UCSC Genome Browser** — The UCSC Genome Browser is an online and downloadable genome browser hosted by the University of California, Santa Cruz (UCSC).
- **Model organism database** — Model organism databases (MODs) are biological databases, or knowledgebases, dedicated to the provision of in-depth biological data for intensively studied model organisms.
- **Genome browser** — The completion of the human genome sequencing in the early 2000s was a turning point in genomics research. Scientists have conducted series of research into the activities of genes and the genome as a whole.
- **Reference genome** — A reference genome is a genome assembly that represents the complete genetic sequence of an organism as a continuous string of nucleotides.
- **GENCODE** — GENCODE is a scientific project in genome research and part of the ENCODE scale-up project.
- **FlyBase** — FlyBase is an online bioinformatics database and the primary repository of genetic and molecular data for the insect family Drosophilidae.
- **Genome Taxonomy Database** — The Genome Taxonomy Database (GTDB) is an online database that maintains information on a proposed nomenclature of prokaryotes, following a phylogenomic approach based on a set of conserved single-copy proteins.

E.2.4 Structure databases

- **Protein Data Bank** — The Protein Data Bank (PDB) is a database for the three-dimensional structural data of large biological molecules such as proteins and nucleic acids, which is overseen by the Worldwide Protein Data Bank (wwPDB).
- **Worldwide Protein Data Bank** — The Worldwide Protein Data Bank (wwPDB) is an organization that maintains the archive of macromolecular structure.
- **EM Data Bank** — The EM Data Bank or Electron Microscopy Data Bank (EMDB) collects 3DEM maps and associated experimental data determined using electron microscopy of biological specimens.

- **AlphaFold** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.
- **CATH database** — The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.
- **Structural Classification of Proteins database** — The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

E.2.5 Function and pathways

- **Gene Ontology** — The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
- **KEGG** — KEGG is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances.
- **Reactome** — Reactome is a free online database of biological pathways. It is manually curated and authored by PhD-level biologists, in collaboration with Reactome editorial staff. The content is cross-referenced to many bioinformatics databases.
- **Ontology (information science)** — In information science, an ontology encompasses a representation, formal naming, and definitions of the categories, properties, and relations between the concepts, data, or entities that pertain to one, many, or all domains of discourse.
- **Biological pathway** — In cell biology, a biological pathway is a series of molecular interactions among molecules within a cell that leads to a specific product or a change in the cell.
- **Gene set enrichment analysis** — Gene set enrichment analysis (GSEA) (also called functional enrichment analysis or pathway enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with different phenotypes (e.g. different organism growth patterns or diseases).
- **Critical Assessment of Function Annotation** — The Critical Assessment of Function Annotation (CAFA) is an ongoing community-driven experiment designed to evaluate computational methods for protein function prediction.
- **Knowledge graph** — In knowledge representation and reasoning, a knowledge graph is a knowledge base that uses a graph-structured data model or topology to represent and operate on data.

E.2.6 Interaction databases

- **STRING** — In molecular biology, STRING is a biological database and web resource of known and predicted protein–protein interactions.
- **BioGRID** — The Biological General Repository for Interaction Datasets (BioGRID) is a curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post-translational modifications created in 2003 (originally referred to as simply the General Repository for Interaction Datasets by Mike Tyers, Bobby-Joe Breitkreutz, and Chris Stark at the Lunenfeld-Tanenbaum Research Institute at Mount Sinai Hospital.
- **Interactome** — In molecular biology, an interactome is the whole set of molecular interactions in a particular cell. The term specifically refers to physical interactions among molecules but can also describe sets of indirect interactions among genes.
- **Protein–protein interaction** — Protein–protein interactions (PPIs) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect.

E.2.7 Expression and proteomics repositories

- **Gene Expression Omnibus** — Gene Expression Omnibus (GEO) is a database for gene expression profiling and DNA methylation profiling managed by the National Center for Biotechnology Information (NCBI).
- **Human Protein Atlas** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.
- **Gene expression profiling** — In the field of molecular biology, gene expression profiling is the measurement of the activity of thousands of genes at once, to create a global picture of cellular function.
- **RNA-Seq** — RNA-Seq is a next-generation sequencing (NGS) technique used to quantify and identify RNA molecules in a biological sample, providing a snapshot of the transcriptome at a specific time.
- **Proteomics** — Proteomics is the large-scale study of proteins. The proteome is the entire set of proteins produced or modified by an organism or system.

E.2.8 Medical and variant data

- **DbSNP** — The Single Nucleotide Polymorphism Database (dbSNP) is a free public archive for genetic variation within and across different species developed and hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI).
- **ClinVar** — ClinVar is a public archive with free access to reports on the relationships between human genetic variations and phenotypes, with supporting evidence. The database includes germline and somatic variants of any size, type or genomic location.
- **Online Mendelian Inheritance in Man** — Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationship.
- **GWAS catalog** — The GWAS Catalog is a free online database that compiles data of genome-wide association studies (GWAS), summarizing unstructured data from different literature sources into accessible high quality data.
- **The Cancer Genome Atlas** — The Cancer Genome Atlas (TCGA) is a project to catalogue the genomic alterations responsible for cancer using genome sequencing and bioinformatics.
- **COSMIC cancer database** — COSMIC is an online database of somatically acquired mutations found in human cancer. Somatic mutations are those that occur in non-germline cells that are not inherited by children.
- **European Genome-phenome Archive** — European Genome-phenome Archive (EGA) is a repository for human biomolecular and phenotypic data in the United Kingdom and Spain.
- **Variant of uncertain significance** — A variant of uncertain significance (VUS) is a genetic variant that has been identified through genetic testing but whose significance to the function or health of an organism is not known.

E.3 Evolution

Wikipedia articles for Chapter 3.

E.3.1 The origin of life

- **Abiogenesis** — Abiogenesis or the origin of life is the natural process by which life arises from non-living matter, such as simple organic compounds.
- **Miller–Urey experiment** — The Miller–Urey experiment, or Miller experiment, was an experiment in chemical synthesis carried out in 1952 that simulated the conditions thought at the time to be present in the atmosphere of the early, prebiotic Earth.
- **RNA world** — The RNA world is a hypothetical stage in the evolutionary history of life on Earth in which self-replicating RNA molecules proliferated before the evolution of DNA and proteins. The term also refers to the hypothesis that posits the existence of this stage.
- **Ribozyme** — Ribozymes are RNA molecules that have the ability to catalyze specific biochemical reactions, including RNA splicing in gene expression, similar to the action of protein enzymes.
- **Iron–sulfur world hypothesis** — The iron–sulfur world hypothesis is a set of proposals for the origin of life and the early evolution of life advanced in a series of articles between 1988 and 1992 by Günter Wächtershäuser, a Munich patent lawyer with a degree in chemistry, who had been encouraged and supported by philosopher Karl R.
- **Protocell** — A protocell is a self-organized, membrane-bound or membraneless compartment that concentrates biomolecules, proposed as a rudimentary precursor to cells during the origin of life. It is a minimal unit that can undergo Darwinian evolution.
- **Earliest known life forms** — The earliest known life forms on Earth may be as old as 4.1 billion years according to biologically fractionated graphite inside a single zircon grain in the Jack Hills range of Australia.
- **Astrobiology** — Astrobiology is a scientific field within the life and environmental sciences that studies the origins, early evolution, distribution, and future of life in the universe by investigating its deterministic conditions and contingent events.

E.3.2 Basic evolutionary concepts

- **Mutation** — In biology, a mutation is an alteration in the nucleic acid sequence of the genome of an organism, virus, or extrachromosomal DNA.
- **Natural selection** — Natural selection is the differential survival and reproduction of individuals due to differences in the relative fitness endowed on them by their own particular complement of observable characteristics.
- **Genetic drift** — Genetic drift is the change in the frequency of an existing allele in a population due to chance.
- **Effective population size** — The effective population size (Ne) is the size of an idealised population that would experience the same rate of genetic drift as the real population.
- **Genetic hitchhiking** — Genetic hitchhiking, also called genetic draft or the hitchhiking effect, is when an allele changes frequency not because it itself is under natural selection, but because it is near another gene that is undergoing a selective sweep and that is on the same DNA chain.
- **Neutral theory of molecular evolution** — The neutral theory of evolution holds that most of the genetic variation within and between species is due to random genetic drift of mutant alleles that are selectively neutral.
- **Nearly neutral theory of molecular evolution** — The nearly neutral theory of molecular evolution is a modification of the neutral theory of molecular evolution that accounts for the fact that not all mutations are either so deleterious such that they can be ignored, or else neutral.
- **Molecular clock** — The molecular clock is a figurative term for a technique that uses the mutation rate of biomolecules to deduce the time in prehistory when two or more life forms diverged.

E.3.3 Homologues, orthologues and gene duplication

- **Homology (biology)** — In biology, homology is similarity in anatomical structures or genes between organisms of different taxa due to shared ancestry, regardless of current functional differences.
- **Sequence homology** — Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life.
- **Gene duplication** — Gene duplication is a mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.
- **Subfunctionalization** — Subfunctionalization was proposed by Stoltzfus (1999) and Force et al.
- **Neofunctionalization** — Neofunctionalization, one of the possible outcomes of functional divergence, occurs when one gene copy, or paralog, takes on a totally new function after a gene duplication event.
- **Gene conversion** — Gene conversion is the process by which one DNA sequence replaces a homologous sequence such that the sequences become identical after the conversion.
- **Concerted evolution** — Concerted evolution is the phenomenon where paralogous genes within one species are more closely related to one another than to members of the same gene family in closely related species.
- **Paleopolyploidy** — Paleopolyploidy is the result of genome duplications which occurred at least several million years ago (MYA). Such an event could either double the genome of a single species (autopolyploidy) or combine those of two species (allopolyploidy).

E.3.4 Protein evolution

- **Protein domain** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure.
- **Tandem repeat** — In genetics, tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other, e.g. ATTCG ATTCG ATTCG, in which the sequence ATTCG is repeated three times.
- **Intrinsically disordered proteins** — In molecular biology, an intrinsically disordered protein (IDP) is a protein that lacks a fixed or ordered three-dimensional structure, typically in the absence of its macromolecular interaction partners, such as other proteins or RNA.
- **Short linear motif** — In molecular biology short linear motifs (SLiMs), linear motifs or minimotifs are short stretches of protein sequence that mediate protein–protein interaction.
- **Direct coupling analysis** — Direct coupling analysis or DCA is an umbrella term comprising several methods for analyzing sequence data in computational biology.
- **Alternative splicing** — Alternative splicing, alternative RNA splicing, or differential splicing is an alternative splicing process during gene expression that allows a single gene to produce different splice variants.
- **Ancestral reconstruction** — Ancestral reconstruction is the extrapolation back in time from measured characteristics of individuals, populations, or species to their common ancestors.

E.3.5 Detecting selection

- **Ka/Ks ratio** — In genetics, the Ka/Ks ratio, also known as ω or dN/dS ratio, is used to estimate the balance between neutral mutations, purifying selection and beneficial mutations acting on a set of homologous protein-coding genes.

- **Models of DNA evolution** — A number of different Markov models of DNA sequence evolution have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution.
- **Negative selection (natural selection)** — In natural selection, negative selection or purifying selection is the selective removal of alleles that are deleterious. This can result in stabilising selection through the purging of deleterious genetic polymorphisms that arise through random mutations.
- **Directional selection** — In population genetics, directional selection is a mode of natural selection in which individuals with a trait at one extreme of a phenotypic distribution have better fitness than individuals with intermediate or opposite extreme phenotypes.
- **McDonald–Kreitman test** — The McDonald–Kreitman test is a statistical test often used by evolutionary and population biologists to detect and measure the amount of adaptive evolution within a species by determining whether adaptive evolution has occurred, and the proportion of substitutions that resulted from positive selection.
- **Tajima's D** — Tajima's D is a population genetic test statistic created by and named after the Japanese researcher Fumio Tajima.
- **Selective sweep** — In genetics, a selective sweep is the process through which a new beneficial mutation that increases its frequency and becomes fixed in a population leads to the reduction or elimination of genetic variation among nucleotide sequences that are near the mutation site.

E.3.6 Genome evolution

- **Horizontal gene transfer** — Horizontal gene transfer (HGT) or lateral gene transfer (LGT) is the movement of genetic material between organisms other than by the ("vertical") transmission of DNA from parent to offspring (reproduction).
- **Bacterial conjugation** — Bacterial conjugation is the transfer of genetic material between bacterial cells by direct cell-to-cell contact or by a bridge-like connection between two cells. This typically takes place through a type IV secretion system, a type of pilus.
- **Transduction (genetics)** — Transduction is the process by which foreign DNA is introduced into a cell by a virus or viral vector. An example is the viral transfer of DNA from one bacterium to another and hence an example of horizontal gene transfer.
- **Genetic transformation** — In molecular biology and genetics, transformation is the genetic alteration of a bacterial cell resulting from the direct uptake and incorporation of exogenous genetic material from its surroundings through the cell membrane(s).
- **Pan-genome** — In the fields of molecular biology and genetics, a pan-genome is the entire set of genes from all strains within a clade. More generally, it is the union of all the genomes of a clade.
- **Gene family** — A gene family is a set of several similar genes, formed by duplication of a single original gene, and generally with similar biochemical functions.
- **Power law** — In mathematics and science, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a relative change in the other quantity proportional to the change raised to a constant exponent: one quantity varies as a power of another.
- **Preferential attachment** — A preferential attachment process is any of a class of processes in which some quantity, typically some form of wealth or credit, is distributed among a number of individuals or objects according to how much they already have, so that those who are already wealthy receive more than those who are not.

E.4 Phylogeny

Wikipedia articles for Chapter 4.

E.4.1 Basic definitions

- **Phylogenetics** — In biology, phylogenetics is the study of the evolutionary history of life using observable characteristics of organisms, which is known as phylogenetic inference.
- **Phylogenetic tree** — A phylogenetic tree or phylogeny is a graphical representation which shows the evolutionary history between a set of species or taxa during a specific time.
- **Cladistics** — Cladistics is an approach to biological classification in which organisms are categorized in groups ("clades") based on hypotheses of most recent common ancestry.
- **Monophyly** — In biological cladistics for the classification of organisms, monophyly is the condition of a taxonomic grouping being a clade – that is, a grouping of organisms which meets these criteria:the grouping contains its own most recent common ancestor, i.e. excludes non-descendants of that common ancestor the grouping contains all the descendants of that common ancestor, without exception
- **Paraphyly** — In taxonomy, a paraphyly is a grouping that consists of the grouping's last common ancestor and some but not all of its descendant lineages. The grouping is said to be paraphyletic with respect to the excluded subgroups.
- **Homology (biology)** — In biology, homology is similarity in anatomical structures or genes between organisms of different taxa due to shared ancestry, regardless of current functional differences.
- **Molecular clock** — The molecular clock is a figurative term for a technique that uses the mutation rate of biomolecules to deduce the time in prehistory when two or more life forms diverged.
- **Newick format** — In mathematics and phylogenetics, Newick tree format is a way of representing graph-theoretical trees with edge lengths using parentheses and commas. It was adopted by James Archie, William H. E. Day, Joseph Felsenstein, Wayne Maddison, Christopher Meacham, F.

E.4.2 Phylogenetic methods

- **Computational phylogenetics** — Computational phylogenetics, phylogeny inference, or phylogenetic inference focuses on computational and optimization algorithms, heuristics, and approaches involved in phylogenetic analyses.
- **UPGMA** — UPGMA is a simple agglomerative (bottom-up) hierarchical clustering method. It also has a weighted variant, WPGMA, and they are generally attributed to Sokal and Michener.
- **Neighbor joining** — In bioinformatics, neighbor joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees, created by Naruya Saitou and Masatoshi Nei in 1987.
- **Maximum parsimony** — In evolutionary biology, maximum parsimony is a method used to build phylogenetic trees based on the principle of simplicity. Under the maximum-parsimony criterion, the best tree will minimize the amount of homoplasy.
- **Long branch attraction** — In phylogenetics, long branch attraction (LBA) is a form of systematic error whereby distantly related lineages are incorrectly inferred to be closely related.
- **Felsenstein's tree-pruning algorithm** — In statistical genetics, Felsenstein's tree-pruning algorithm, attributed to Joseph Felsenstein, is an algorithm for efficiently computing the likelihood of an evolutionary tree from nucleic acid sequence data.
- **Bayesian inference in phylogeny** — Bayesian inference of phylogeny combines the information in the prior and in the data likelihood to create the so-called posterior probability of trees, which is the probability that the tree is correct given the data, the prior and the likelihood model.
- **Models of DNA evolution** — A number of different Markov models of DNA sequence evolution have been proposed. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution.

E.4.3 Alignments, genomes and conflicting trees

- **Phylogenomics** — Phylogenomics is the intersection of the fields of evolution and genomics. The term has been used in multiple ways to refer to analysis that involves genome data and evolutionary reconstructions.
- **Multiple sequence alignment** — Multiple sequence alignment (MSA) is the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.
- **Incomplete lineage sorting** — Incomplete lineage sorting (ILS) (also referred to as hemiplasy, deep coalescence, retention of ancestral polymorphism, or trans-species polymorphism) is a phenomenon in evolutionary biology and population genetics that results in discordance between species and gene trees.
- **Multispecies coalescent process** — Multispecies Coalescent Process is a stochastic process model that describes the genealogical relationships for a sample of DNA sequences taken from several species. It represents the application of coalescent theory to the case of multiple species.
- **Coalescent theory** — Coalescent theory is a model of how alleles sampled from a population may have originated from a common ancestor.
- **Phylogenetic reconciliation** — In phylogenetics, reconciliation is an approach to connect the history of two or more coevolving biological entities.
- **Phylogenetic network** — A phylogenetic network is any graph used to visualize evolutionary relationships between nucleotide sequences, genes, chromosomes, genomes, or species.
- **Horizontal gene transfer** — Horizontal gene transfer (HGT) or lateral gene transfer (LGT) is the movement of genetic material between organisms other than by the ("vertical") transmission of DNA from parent to offspring (reproduction).

E.4.4 The tree of life

- **Tree of life (biology)** — The tree of life or universal tree of life is a metaphor, conceptual model, and research tool used to explore the evolution of life and describe the relationships between organisms, both living and extinct, as described in a famous passage in Charles Darwin's On the Origin of Species (1859).The affinities of all the beings of the same class have sometimes been represented by a great tree.
- **Three-domain system** — The three-domain system is a taxonomic classification system that groups all cellular life into three domains, namely Archaea, Bacteria and Eukarya, introduced by Carl Woese, Otto Kandler and Mark Wheelis in 1990.
- **Two-domain system** — The two-domain system is a biological classification system that groups all cellular organisms in the tree of life into two domains: Archaea and Bacteria, while eukaryotes are considered a specialized clade under Archaea.
- **Last universal common ancestor** — The last universal common ancestor (LUCA) is the hypothesized latest common ancestral population of single-celled organisms from which all subsequent life forms descend under the three-domain system of Bacteria, Archaea, and Eukarya.
- **Asgard (archaea)** — Asgard archaea, or Promethearchaeati, are a kingdom belonging to the domain Archaea that contain eukaryotic signature proteins.
- **Eukaryogenesis** — Eukaryogenesis, the process which created the eukaryotic cell and lineage, is a milestone in the evolution of life, since eukaryotes include all complex cells and almost all multicellular organisms.
- **Hominidae** — The Hominidae, whose members are known as the great apes, are a taxonomic family of primates that includes eight extant species in four genera: Pongo ; Gorilla ; Pan ; and Homo, of which only modern humans remain.
- **Human evolution** — Homo sapiens is a distinct species of the hominid family of primates, which includes all the great apes. Over their evolutionary history, humans gradually developed traits such as bipedalism, dexterity, and complex language.

E.5 Genomics

Wikipedia articles for Chapter 5.

E.5.1 Sequencing methods

- **DNA sequencing** — DNA sequencing is the process of determining the nucleic acid sequence – the order of nucleotides in DNA. It includes any method or technology that is used to determine the order of the four bases: adenine, thymine, cytosine, and guanine.
- **Sanger sequencing** — Sanger sequencing is a method of DNA sequencing that involves electrophoresis and is based on the random incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.
- **Massive parallel sequencing** — Massively parallel sequencing (MPS) is any of several high-throughput approaches to DNA sequencing using the concept of massively parallel processing; it is also called next-generation sequencing (NGS) or second-generation sequencing.
- **Illumina dye sequencing** — Illumina dye sequencing is a technique used to determine the series of base pairs in DNA, also known as DNA sequencing. The reversible terminated chemistry concept was invented by Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris.
- **Single-molecule real-time sequencing** — Single-molecule real-time (SMRT) sequencing is a parallelized single molecule DNA sequencing method. Single-molecule real-time sequencing utilizes a zero-mode waveguide (ZMW).
- **Nanopore sequencing** — Nanopore sequencing is a third generation approach used in the sequencing of biopolymers — specifically, polynucleotides in the form of DNA or RNA.
- **Phred quality score** — A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing.
- **Coverage (genetics)** — In genetics, coverage is one of several measures of the depth or completeness of DNA sequencing, and is more specifically expressed in any of the following terms:Sequence coverage is the number of unique reads that include a given nucleotide in the reconstructed sequence.

E.5.2 Genome assembly

- **Sequence assembly** — In bioinformatics, sequence assembly refers to aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence.
- **De novo sequence assemblers** — De novo sequence assemblers are a type of program that assembles short nucleotide sequences into longer ones without the use of a reference genome. These are most commonly used in bioinformatic studies to assemble genomes or transcriptomes.
- **De Bruijn graph** — In graph theory, an n-dimensional De Bruijn graph of m symbols is a directed graph representing overlaps between sequences of symbols.
- **Eulerian path** — In graph theory, an Eulerian trail is a trail in a finite graph that visits every edge exactly once. Similarly, an Eulerian circuit or Eulerian cycle is an Eulerian trail that starts and ends on the same vertex.
- **N50, L50, and related statistics** — In computational biology, N50 and L50 are statistics of a set of contig or scaffold lengths. The N50 is similar to a mean or median of lengths, but has greater weight given to the longer contigs.
- **Scaffolding (bioinformatics)** — Scaffolding is a technique used in bioinformatics. It is defined as follows: Link together a non-contiguous series of genomic sequences into a scaffold, consisting of sequences separated by gaps of known length.

- **Hybrid genome assembly** — In bioinformatics, hybrid genome assembly refers to utilizing various sequencing technologies to achieve the task of assembling a genome from fragmented, sequenced DNA resulting from shotgun sequencing.
- **Whole genome sequencing** — Whole genome sequencing (WGS), also known as full genome sequencing or just genome sequencing, is the process of determining the entirety of the DNA sequence of an organism's genome at a single time.

E.5.3 Mapping and variant calling

- **Sequence alignment** — In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.
- **Burrows–Wheeler transform** — The Burrows–Wheeler transform (BWT) rearranges a character string into runs of similar characters, in a manner that can be reversed to recover the original string.
- **FASTQ format** — FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.
- **SAM (file format)** — Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence developed by Heng Li and Bob Handsaker et al.
- **SNV calling from NGS data** — SNV calling from NGS data is any of a range of methods for identifying the existence of single nucleotide variants (SNVs) from the results of next generation sequencing (NGS) experiments.
- **Variant Call Format** — The Variant Call Format or VCF is a standard text file format used in bioinformatics for storing gene sequence or DNA sequence variations.
- **Structural variation** — Genomic structural variation is the variation in structure of an organism's chromosome, such as deletions, duplications, copy-number variants, insertions, inversions and translocations.

E.5.4 Gene prediction

- **Gene prediction** — In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes.
- **Open reading frame** — In molecular biology and genetics, reading frames are defined as spans of DNA sequence between the start and stop codons.
- **Hidden Markov model** — In probability theory, a hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process. An HMM requires that there be an observable process whose outcomes depend on the outcomes of in a known way.
- **RNA splicing** — RNA splicing is a process in molecular biology where a newly-made precursor messenger RNA (pre-mRNA) transcript is transformed into a mature messenger RNA (mRNA). It works by removing all the introns and splicing back together exons.
- **GENSCAN** — In bioinformatics, GENSCAN is a program to identify complete gene structures in genomic DNA. It is a GHMM-based program that can be used to predict the location of genes and their exon-intron boundaries in genomic sequences from a variety of organisms.
- **GeneMark** — GeneMark is a generic name for a family of ab initio gene prediction algorithms and software programs developed at the Georgia Institute of Technology in Atlanta.
- **Genome project** — Genome projects are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism and to annotate protein-coding genes and other important genome-encoded features.

E.5.5 Metagenomics

- **Metagenomics** — Metagenomics is the study of all genetic material from all organisms in a particular environment, providing insights into their composition, diversity, and functional potential.
- **16S ribosomal RNA** — 16S ribosomal RNA is the RNA component of the 30S subunit of a prokaryotic ribosome. It binds to the Shine-Dalgarno sequence and provides most of the SSU structure.
- **Operational taxonomic unit** — An operational taxonomic unit (OTU) is an operational definition used to classify groups of closely related individuals. The term was originally introduced in 1963 by Robert R. Sokal and Peter H. A.
- **Amplicon sequence variant** — An amplicon sequence variant (ASV) is any one of the inferred single DNA sequences recovered from a high-throughput analysis of marker genes.
- **Binning (metagenomics)** — In metagenomics, binning is the computational process of grouping assembled contigs and assigning them to their separate genomes of origin.
- **Metagenome-assembled genome** — Metagenomics is the study of all genetic material from all organisms in a particular environment, providing insights into their composition, diversity, and functional potential.

E.5.6 Regulatory and epigenomic sequencing data

- **ChIP sequencing** — ChIP-sequencing, also known as ChIP-seq, is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.
- **ATAC-seq** — ATAC-seq is a laboratory technique used in molecular biology to assess genome-wide chromatin accessibility.
- **Bisulfite sequencing** — Bisulfite sequencing (also known as bisulphite sequencing) is the use of bisulfite treatment of DNA before routine sequencing to determine the pattern of methylation. DNA methylation was the first discovered epigenetic mark, and remains the most studied.
- **DNA methylation** — DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence.
- **Cross-linking immunoprecipitation** — Cross-linking and immunoprecipitation is a method used in molecular biology that combines UV crosslinking with immunoprecipitation in order to identify RNA binding sites of proteins on a transcriptome-wide scale, thereby increasing our understanding of post-transcriptional regulatory networks.

E.5.7 Human genetics: finding disease genes and variants

- **Genome-wide association study** — In genomics, a genome-wide association study, is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait.
- **Linkage disequilibrium** — Linkage disequilibrium, often abbreviated to LD, is a term in population genetics referring to the association of genes, usually linked genes, in a population. It has become an important tool in medical genetics and other fields.
- **Polygenic score** — In genetics, a polygenic score (PGS) is a number that summarizes the estimated effect of many genetic variants on an individual's phenotype.
- **Missing heritability problem** — In genetics, the missing heritability problem refers to a difference between heritability estimates obtained from early genome-wide association studies (GWAS)

and heritability estimates from twin and family data across many physical and mental traits, including diseases, behaviors, and other phenotypes.
- **Exome sequencing** — Exome sequencing, also known as whole exome sequencing (WES), is a genomic technique for sequencing all of the protein-coding regions of genes in a genome. It consists of two steps: the first step is to select only the subset of DNA that encodes proteins.
- **Genetic linkage** — Genetic linkage is the tendency of DNA sequences that are close together on a chromosome to be inherited together during the meiosis phase of sexual reproduction.

E.6 Genomes

Wikipedia articles for Chapter 6.

E.6.1 Background: genes, genomes and cells

- **DNA** — Deoxyribonucleic acid is a polymer composed of two polynucleotide chains that coil around each other to form a double helix.
- **RNA** — Ribonucleic acid (RNA) is a polymeric molecule that is essential for most biological functions, either by performing the function itself or by forming a template for the production of proteins. RNA and deoxyribonucleic acid (DNA) are nucleic acids.
- **Gene** — In biology, the word gene has two meanings. The Mendelian gene is a basic unit of heredity. The molecular gene is a sequence of nucleotides in DNA that is transcribed to produce RNA.
- **Central dogma of molecular biology** — The central dogma of molecular biology deals with the flow of genetic information within a biological system. It is often stated as "DNA makes RNA, and RNA makes protein", although this is not its original meaning.
- **Transcription (biology)** — Transcription is the process of duplicating a segment of DNA into RNA for the purpose of gene expression. Some segments of DNA are transcribed into RNA molecules that can encode proteins, called messenger RNA (mRNA).
- **Translation (biology)** — Translation is the process in biological cells in which proteins are produced using RNA molecules as templates. The generated protein is a sequence of amino acids determined by the sequence of nucleotides in the RNA.
- **Genetic code** — Genetic code is a set of rules used by living cells to translate information encoded within genetic material into proteins.
- **Model organism** — A model organism is a non-human species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the model organism will provide insight into the workings of other organisms.

E.6.2 From the first genome projects

- **Genome project** — Genome projects are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism and to annotate protein-coding genes and other important genome-encoded features.
- **Human Genome Project** — The Human Genome Project (HGP) was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying, mapping and sequencing all of the genes of the human genome from both a physical and a functional standpoint.
- **Shotgun sequencing** — In genetics, shotgun sequencing is a method used for sequencing random DNA strands. It is named by analogy with the rapidly expanding, quasi-random shot grouping of a shotgun.

- **Bermuda Principles** — The Bermuda Principles set out rules for the rapid and public release of DNA sequence data.
- **Genome size** — Genome size is the total amount of DNA contained within one copy of a single complete genome. It is typically measured in terms of mass in picograms or less frequently in daltons, or as the total number of nucleotide base pairs, usually in megabases.
- **C-value** — C-value is the amount, in picograms, of DNA contained within a haploid nucleus or one half the amount in a diploid somatic cell of a eukaryotic organism.
- **Polyploidy** — Polyploidy is a condition in which the cells of an organism have more than two paired sets of (homologous) chromosomes.

E.6.3 Differences across the tree of life

- **Prokaryote** — A prokaryote, less commonly spelled procaryote, is a cellular organism that lacks a distinct cell nucleus or other membrane-bound organelles.
- **Bacterial genome** — Bacterial genomes are generally smaller and less varied in size between species when compared with genomes of eukaryotes. Bacterial genomes can range in size anywhere from about 130 kbp to over 14 Mbp.
- **Operon** — In genetics, an operon is a functioning unit of DNA containing a cluster of genes under the control of a single promoter.
- **GC skew** — GC skew is when the nucleotides guanine and cytosine are over- or under-abundant in a particular region of DNA or RNA. GC skew is also a statistical method for measuring strand-specific guanine overrepresentation.
- **Intron** — An intron is any nucleotide sequence within a gene that is not expressed or operative in the final RNA product. The word intron is derived from the term intragenic region, i.e., a region inside a gene.
- **Paleopolyploidy** — Paleopolyploidy is the result of genome duplications which occurred at least several million years ago (MYA). Such an event could either double the genome of a single species (autopolyploidy) or combine those of two species (allopolyploidy).
- **Symbiogenesis** — Symbiogenesis, also known as the endosymbiotic theory or serial endosymbiotic theory,) is the leading scientific theory for the evolutionary origin of eukaryotes from prokaryotic ancestors.
- **Mitochondrial DNA** — Mitochondrial DNA is the DNA located in the mitochondria organelles in a eukaryotic cell that converts chemical energy from organic compounds into adenosine triphosphate (ATP).

E.6.4 Genomes in populations

- **Pan-genome** — In the fields of molecular biology and genetics, a pan-genome is the entire set of genes from all strains within a clade. More generally, it is the union of all the genomes of a clade.
- **Genetic drift** — Genetic drift is the change in the frequency of an existing allele in a population due to chance.
- **Effective population size** — The effective population size (Ne) is the size of an idealised population that would experience the same rate of genetic drift as the real population.
- **Drift-barrier hypothesis** — The drift-barrier hypothesis is an evolutionary hypothesis formulated by Michael Lynch in 2010 as an application of mutation-selection balance to mutation rates.
- **Recent African origin of modern humans** — The recent African origin of modern humans or the "Out of Africa" theory (OOA) holds that present-day humans outside Africa descend mainly from a single expansion of anatomically modern humans from Africa about 70,000–50,000 years ago.

- **Neanderthal genome project** — The Neanderthal genome project is an effort, founded in July 2006, of a group of scientists to sequence the Neanderthal genome.
- **Interbreeding between archaic and modern humans** — Interbreeding between archaic humans and anatomically modern humans took place during the Middle Paleolithic and early Upper Paleolithic.
- **Denisovan** — The Denisovans or Denisova hominins are an extinct species or subspecies of archaic human that ranged across Asia during the Middle to Late Pleistocene, approximately 200,000–32,000 years ago. Most of what is known about Denisovans comes from DNA evidence.

E.6.5 Repeats and the 3D genome

- **Transposable element** — Transposable elements (TEs), also known as transposons, jumping genes, or mobile genetic elements, are DNA sequences that can change their position, or translocate, within a genome.
- **Retrotransposon** — Retrotransposons are mobile elements which move in the host genome by converting their transcribed RNA into DNA through reverse transcription.
- **Alu element** — An Alu element is a short stretch of DNA originally characterized by the action of the Arthrobacter luteus (Alu) restriction endonuclease. Alu elements are the most abundant transposable elements in the human genome, present in excess of one million copies.
- **Selfish genetic element** — Selfish genetic elements are genetic segments that can enhance their own transmission at the expense of other genes in the genome, even if this has no positive or a net negative effect on organismal fitness.
- **Tandem repeat** — In genetics, tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other, e.g. ATTCG ATTCG ATTCG, in which the sequence ATTCG is repeated three times.
- **Satellite DNA** — Satellite DNA, also known as highly repetitive DNA, consists of very large arrays of tandemly repeating, non-coding DNA. Satellite DNA is the main component of functional centromeres, and form the main structural constituent of heterochromatin.
- **Segmental duplication** — Low copy repeats (LCRs), also known as segmental duplications (SDs), or duplicons, are DNA sequences with high levels of sequence identity, present in multiple locations within a genome.
- **Chromatin** — Chromatin is a complex of DNA and protein responsible for condensing and packaging chromosomal DNA. Chromatin is found in both bacterial and eukaryotic cells.
- **Chromosome conformation capture** — Chromosome conformation capture techniques are a set of molecular biology methods used to analyze the spatial organization of chromatin in a cell.
- **Topologically associating domain** — A topologically associating domain (TAD) is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD.

E.6.6 RNA genes and regulatory elements

- **Non-coding RNA** — A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein. The DNA sequence from which a functional non-coding RNA is transcribed is often called a non-coding gene or an RNA gene.
- **MicroRNA** — Micro ribonucleic acid are small, single-stranded, non-coding RNA molecules containing 21–23 nucleotides.
- **Long non-coding RNA** — Long non-coding RNAs are a type of RNA, generally defined as transcripts of more than 200 nucleotides that are not translated into protein.

- **Enhancer (genetics)** — In genetics, an enhancer is a short region of DNA that can be bound by proteins (activators) to increase the likelihood that transcription of a particular gene will occur. These proteins are usually referred to as transcription factors.
- **Non-coding DNA** — Non-coding DNA (ncDNA) sequences are components of an organism's DNA that do not encode protein sequences. Some non-coding DNA is transcribed into functional non-coding RNA molecules.
- **Junk DNA** — Junk DNA is a DNA sequence that has no known biological function.
- **ENCODE** — The Encyclopedia of DNA Elements (ENCODE) is a public research project which aims "to build a comprehensive parts list of functional elements in the human genome."

E.6.7 Gene order, synteny and genome rearrangements

- **Synteny** — In genetics, the term synteny refers to two related concepts:In classical genetics, synteny describes the physical co-localization of genetic loci on the same chromosome within an individual or species.
- **Chromosomal rearrangement** — In genetics, a chromosomal rearrangement is a mutation that is a type of chromosome abnormality involving a change in the structure of the native chromosome.
- **Dot plot (bioinformatics)** — In bioinformatics a dot plot is a graphical method for comparing two biological sequences and identifying regions of close similarity after sequence alignment. It is a type of recurrence plot.
- **Comparative genomics** — Comparative genomics is a branch of biological research that examines genome sequences across a spectrum of species, spanning from humans and mice to a diverse array of organisms from bacteria to chimpanzees.
- **Chromosomal inversion** — An inversion is a chromosome rearrangement in which a segment of a chromosome becomes inverted within its original position.

E.6.8 Pseudogenes, new genes and small proteins

- **Pseudogene** — Pseudogenes are nonfunctional segments of DNA that resemble functional genes. Pseudogenes can be formed from both protein-coding genes and non-coding genes.
- **De novo gene birth** — De novo gene birth is the process by which new genes evolve from non-coding DNA. De novo genes represent a subset of novel genes, and may be protein-coding or instead act as RNA genes.
- **Orphan gene** — Orphan genes, ORFans, or taxonomically restricted genes are genes that lack a detectable homologue outside of a given species or lineage. Most genes have known homologues.
- **Small open reading frame** — In molecular biology and genetics, reading frames are defined as spans of DNA sequence between the start and stop codons.
- **Retrogene** — Gene duplication is a mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.

E.7 Expression analysis

Wikipedia articles for Chapter 7.

E.7.1 The first genome-wide methods: microarrays

- **DNA microarray** — A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome.
- **Gene expression profiling** — In the field of molecular biology, gene expression profiling is the measurement of the activity of thousands of genes at once, to create a global picture of cellular function.
- **Serial analysis of gene expression** — Serial Analysis of Gene Expression (SAGE) is a transcriptomic technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest in the form of small tags that correspond to fragments of those transcripts.
- **Quantile normalization** — In statistics, quantile normalization is a technique for making two distributions identical in statistical properties.
- **MA plot** — Within computational biology, an MA plot is an application of a Bland–Altman plot for visual representation of genomic data.
- **Affymetrix** — Affymetrix is now Applied Biosystems, a brand of DNA microarray products sold by Thermo Fisher Scientific that originated with an American biotechnology research and development and manufacturing company of the same name.
- **Microarray analysis techniques** — Microarray analysis techniques are used in interpreting the data generated from experiments on DNA, RNA, and protein microarrays, which allow researchers to investigate the expression state of a large number of genes – in many cases, an organism's entire genome – in a single experiment.

E.7.2 RNA sequencing

- **RNA-Seq** — RNA-Seq is a next-generation sequencing (NGS) technique used to quantify and identify RNA molecules in a biological sample, providing a snapshot of the transcriptome at a specific time.
- **Transcriptomics technologies** — Transcriptomics technologies are the techniques used to study an organism's transcriptome, the sum of all of its RNA transcripts. The information content of an organism is recorded in the DNA of its genome and expressed through transcription.
- **Alternative splicing** — Alternative splicing, alternative RNA splicing, or differential splicing is an alternative splicing process during gene expression that allows a single gene to produce different splice variants.
- **Expectation–maximization algorithm** — In statistics, an expectation–maximization (EM) algorithm is an iterative method to find (local) maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.
- **Batch effect** — In molecular biology, a batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment.

E.7.3 Testing and statistics

- **Negative binomial distribution** — In probability theory and statistics, the negative binomial distribution, also called a Pascal distribution, is a discrete probability distribution that models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified/constant/fixed number of successes occur.
- **Generalized linear model** — In statistics, a generalized linear model (GLM) is a flexible generalization of ordinary linear regression.

- **P-value** — In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.
- **Multiple comparisons problem** — Multiple comparisons, multiplicity or multiple testing problem occurs when many statistical tests are performed on the same dataset.
- **False discovery rate** — In statistics, the false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons.
- **Hierarchical clustering** — In data mining and statistics, hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters.
- **Principal component analysis** — Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.
- **T-distributed stochastic neighbor embedding** — t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.

E.7.4 From gene lists to biology

- **Gene set enrichment analysis** — Gene set enrichment analysis (GSEA) (also called functional enrichment analysis or pathway enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with different phenotypes (e.g. different organism growth patterns or diseases).
- **Gene Ontology** — The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
- **Hypergeometric distribution** — In probability theory and statistics, the hypergeometric distribution is a discrete probability distribution that describes the probability of successes in draws, without replacement, from a finite population of size that contains exactly objects with that feature, where in each draw is either a success or a failure.
- **Gene co-expression network** — A gene co-expression network (GCN) is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them.
- **Weighted correlation network analysis** — Weighted correlation network analysis, also known as weighted gene co-expression network analysis (WGCNA), is a widely used data mining method especially for studying biological networks based on pairwise correlations between variables.
- **Expression quantitative trait loci** — An expression quantitative trait locus (eQTL) is a type of quantitative trait locus (QTL), a genomic locus that is associated with phenotypic variation for a specific, quantifiable trait.
- **Human Protein Atlas** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

E.7.5 Single-cell genomics

- **Single-cell transcriptomics** — Single-cell transcriptomics refers to the quantification and analysis of the transcriptomes of individual cells.
- **Single-cell sequencing** — Single-cell sequencing examines the nucleic acid sequence information from individual cells with optimized next-generation sequencing technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.

- **Unique molecular identifier** — Unique molecular identifiers (UMIs), or molecular barcodes (MBC) are short sequences or molecular "tags" added to DNA fragments in some next generation sequencing library preparation protocols to identify the input DNA molecule.
- **CITE-Seq** — CITE-Seq is a method for performing RNA sequencing along with gaining quantitative and qualitative information on surface proteins with available antibodies on a single cell level.
- **RNA velocity** — RNA velocity is a way of measuring the changes in gene expression. It is derived from ratio of spliced to unspliced RNA in the cell over time. Monitoring this value over time allows to see if the gene of interest is induced or repressed.
- **Spatial transcriptomics** — Spatial transcriptomics, or spatially resolved transcriptomics, is a method that captures positional context of transcriptional activity within intact tissue.
- **Foundation model** — In artificial intelligence, a foundation model (FM), also known as large x model, is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases.
- **Human Cell Atlas** — The Human Cell Atlas is a global project to describe all cell types in the human body, to understand human health and for diagnosing, monitoring, and treating disease.

E.7.6 Single-cell sequencing and cancer evolution

- **Somatic evolution in cancer** — Somatic evolution is the accumulation of mutations and epimutations in somatic cells during a lifetime, and the effects of those mutations and epimutations on the fitness of those cells.
- **Tumour heterogeneity** — Tumour heterogeneity describes the observation that different tumour cells can show distinct morphological and phenotypic profiles, including cellular morphology, gene expression, metabolism, motility, proliferation, and metastatic potential.
- **Cancer stem cell** — Cancer stem cells (CSCs) are cancer cells that possess characteristics associated with normal stem cells, specifically the ability to give rise to all cell types found in a particular cancer sample.
- **Single-cell sequencing** — Single-cell sequencing examines the nucleic acid sequence information from individual cells with optimized next-generation sequencing technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.
- **Copy number variation** — Copy number variation (CNV) is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals.
- **Phylogenetic tree** — A phylogenetic tree or phylogeny is a graphical representation which shows the evolutionary history between a set of species or taxa during a specific time.
- **Infinite sites model** — The Infinite sites model (ISM) is a mathematical model of molecular evolution first proposed by Motoo Kimura in 1969. Like other mutation models, the ISM provides a basis for understanding how mutation develops new alleles in DNA sequences.

E.8 Proteomics

Wikipedia articles for Chapter 8.

E.8.1 Mass spectrometry methods

- **Proteomics** — Proteomics is the large-scale study of proteins. The proteome is the entire set of proteins produced or modified by an organism or system.

- **Mass spectrometry** — Mass spectrometry (MS) is an analytical technique that is used to measure the mass-to-charge ratio of ions. The results are presented as a mass spectrum, a plot of intensity as a function of the mass-to-charge ratio.
- **Liquid chromatography–mass spectrometry** — Liquid chromatography–mass spectrometry (LC–MS) is an analytical chemistry technique that combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of mass spectrometry (MS).
- **Electrospray ionization** — Electrospray ionization (ESI) is a technique used in mass spectrometry to produce ions using an electrospray in which a high voltage is applied to a liquid to create an aerosol.
- **Matrix-assisted laser desorption/ionization** — In mass spectrometry, matrix-assisted laser desorption/ionization (MALDI) is an ionization technique that uses a laser energy-absorbing matrix to create ions from large molecules with minimal fragmentation.
- **Orbitrap** — In mass spectrometry, Orbitrap is an ion trap mass analyzer consisting of an outer barrel-like electrode and a coaxial inner spindle-like electrode that traps ions in an orbital motion around the spindle.
- **Tandem mass spectrometry** — Tandem mass spectrometry, also known as MS/MS or MS2, is a technique in instrumental analysis where two or more stages of analysis using one or more mass analyzers are performed with an additional reaction step in between these analyses to increase their abilities to analyse chemical samples.
- **Collision-induced dissociation** — Collision-induced dissociation (CID), also known as collisionally activated dissociation (CAD), is a mass spectrometry technique to induce fragmentation of selected ions in the gas phase.
- **Bottom-up proteomics** — Bottom-up proteomics is a common method to identify proteins and characterize their amino acid sequences and post-translational modifications by proteolytic digestion of proteins prior to analysis by mass spectrometry.
- **Top-down proteomics** — Top-down proteomics is a method of protein identification capable of identifying and quantitating unique proteoforms through the analysis of intact proteins. The name is derived from the similar approach to DNA sequencing.
- **Data-independent acquisition** — In mass spectrometry, data-independent acquisition (DIA) is a method of molecular structure determination in which all ions within a selected m/z range are fragmented and analyzed in a second stage of tandem mass spectrometry.
- **Stable isotope labeling by amino acids in cell culture** — Stable isotope labeling by amino acids in cell culture (SILAC) is a technique based on mass spectrometry that detects differences in protein abundance among samples using non-radioactive isotopic labeling. It is a popular method for quantitative proteomics.
- **Isobaric labeling** — Isobaric labeling is a mass spectrometry strategy used in quantitative proteomics. Peptides or proteins are labeled with chemical groups that have nominally identical mass (isobaric), but vary in terms of distribution of heavy isotopes in their structure.
- **Label-free quantification** — Label-free quantification is a method in mass spectrometry that aims to determine the relative amount of proteins in two or more biological samples.

E.8.2 Algorithms for peptide identification

- **Shotgun proteomics** — Shotgun proteomics refers to the use of bottom-up proteomics techniques in identifying proteins in complex mixtures using a combination of high performance liquid chromatography combined with mass spectrometry.
- **Tandem mass spectrometry** — Tandem mass spectrometry, also known as MS/MS or MS2, is a technique in instrumental analysis where two or more stages of analysis using one or more mass analyzers are performed with an additional reaction step in between these analyses to increase their abilities to analyse chemical samples.

- **Peptide mass fingerprinting** — Peptide mass fingerprinting (PMF), also known as protein fingerprinting, is an analytical technique for protein identification in which the unknown protein of interest is first cleaved into smaller peptides, whose absolute masses can be accurately measured with a mass spectrometer such as MALDI-TOF or ESI-TOF.
- **Mascot (software)** — Mascot is a software search engine that uses mass spectrometry data to identify proteins from peptide sequence databases. Mascot is widely used by research facilities around the world.
- **De novo peptide sequencing** — In mass spectrometry, de novo peptide sequencing is the method in which a peptide amino acid sequence is determined from tandem mass spectrometry.
- **Proteogenomics** — Proteogenomics is a field of biological research that utilizes a combination of proteomics, genomics, and transcriptomics to aid in the discovery and identification of peptides.
- **Edman degradation** — Edman degradation, developed by Pehr Edman, is a method of sequencing amino acids in a peptide. In this method, the amino-terminal residue is labeled and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues.

E.8.3 Statistics and error control

- **False discovery rate** — In statistics, the false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons.
- **Q-value (statistics)** — In statistical hypothesis testing, specifically multiple hypothesis testing, the q-value in the Storey procedure provides a means to estimate the positive false discovery rate (pFDR).
- **Multiple comparisons problem** — Multiple comparisons, multiplicity or multiple testing problem occurs when many statistical tests are performed on the same dataset.
- **Weak supervision** — Weak supervision is a paradigm in machine learning, the relevance and notability of which increased with the advent of large language models due to the large amount of data required to train them.
- **Support vector machine** — In machine learning, a support vector machine (SVM) or support vector network is a supervised max-margin model with associated learning algorithms that analyze data for classification and regression analysis.
- **Shotgun proteomics** — Shotgun proteomics refers to the use of bottom-up proteomics techniques in identifying proteins in complex mixtures using a combination of high performance liquid chromatography combined with mass spectrometry.
- **Label-free quantification** — Label-free quantification is a method in mass spectrometry that aims to determine the relative amount of proteins in two or more biological samples.

E.8.4 Proteomics beyond protein abundance

- **Post-translational modification** — Post-translational modifications (PTMs) are the covalent processes of changing proteins following their synthesis, and release from ribosomes.
- **Phosphoproteomics** — Phosphoproteomics is a branch of proteomics that identifies, catalogs, and characterizes proteins containing a phosphate group as a posttranslational modification.
- **Protein–protein interaction** — Protein–protein interactions (PPIs) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect.
- **Tandem affinity purification** — Tandem affinity purification (TAP) is an immunoprecipitation-based purification technique for studying protein–protein interactions. The goal is to extract from a cell only the protein of interest, in complex with any other proteins it interacted with.

- **Proximity labeling** — Enzyme-catalyzed proximity labeling (PL), also known as proximity-based labeling, is a laboratory technique that labels biomolecules, usually proteins or RNA, proximal to a protein of interest.
- **Hydrogen–deuterium exchange** — Hydrogen–deuterium exchange is a chemical reaction in which a covalently bonded hydrogen atom is replaced by a deuterium atom, or vice versa.
- **Thermal shift assay** — A thermal shift assay (TSA) measures the stability of a protein by the change in its thermal denaturation temperature. This is done under varying conditions, such as variations in drug concentration, buffer formulation, redox potential, or sequence mutation.
- **Mass spectrometry imaging** — Mass spectrometry imaging (MSI) is a technique used in mass spectrometry to visualize the spatial distribution of molecules, as biomarkers, metabolites, peptides or proteins by their molecular masses.

E.8.5 Proteomics without mass spectrometry

- **Human Protein Atlas** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.
- **Immunohistochemistry** — Immunohistochemistry is a form of immunostaining. It involves the process of selectively identifying antigens in cells and tissue, by exploiting the principle of antibodies binding specifically to antigens in biological tissues.
- **Antibody** — An antibody (Ab), or immunoglobulin (Ig), is a large protein belonging to the immunoglobulin superfamily which is used by the immune system to identify and neutralize antigens such as those that exist on bacteria and virus cells, including those that cause disease.
- **Proximity extension assay** — The proximity extension assay (PEA) is a method for detecting and quantifying the amount of many specific proteins present in a biological sample such as serum or plasma.
- **Aptamer** — Aptamers are oligomers of artificial ssDNA, RNA, XNA, or peptide that bind a specific target molecule, or family of target molecules.
- **Mass cytometry** — Mass cytometry is a high-dimensional single-cell analysis technique that integrates flow cytometry with time of flight mass spectrometry. It is used for the determination of the properties of cells (cytometry).
- **Protein sequencing** — Protein sequencing is the practical process of determining the amino acid sequence of all or part of a protein or peptide. This may serve to identify the protein or characterize its post-translational modifications.
- **Nanopore sequencing** — Nanopore sequencing is a third generation approach used in the sequencing of biopolymers — specifically, polynucleotides in the form of DNA or RNA.

E.8.6 Repositories and standards

- **Proteomics Identifications Database** — The PRIDE is a public data repository of mass spectrometry-based proteomics data, and is maintained by the European Bioinformatics Institute as part of the Proteomics Team.
- **PeptideAtlas** — PeptideAtlas is a proteomics data resource that gathers tandem mass spectrometry datasets from around the world, reprocesses them with the Trans-Proteomic Pipeline, and makes the combined result freely available to the community.
- **Proteomics Standards Initiative** — The Proteomics Standards Initiative (PSI) is a working group of the Human Proteome Organization. It aims to define data standards for proteomics to facilitate data comparison, exchange and verification.

- **Mass spectrometry data format** — Mass spectrometry is a scientific technique for measuring the mass-to-charge ratio of ions.
- **Human Proteome Organization** — The Human Proteome Organization (HUPO) is an international consortium of national proteomics research associations, government researchers, academic institutions, and industry partners.
- **Human Proteome Project** — The Human Proteome Project (HPP) is a collaborative effort coordinated by the Human Proteome Organization. Its stated goal is to experimentally observe all of the proteins produced by the sequences translated from the human genome.

E.9 Protein structure

Wikipedia articles for Chapter 9.

E.9.1 CASP: a blind test of structure prediction

- **CASP** — Critical Assessment of Structure Prediction (CASP), sometimes called Critical Assessment of Protein Structure Prediction, is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994.
- **Global distance test** — The global distance test (GDT), also written as GDT_TS to represent "total score", is a measure of similarity between two protein structures with known amino acid correspondences but different tertiary structures.
- **Template modeling score** — In bioinformatics, the template modeling score or TM-score is a measure of similarity between two protein structures.
- **Root-mean-square deviation of atomic positions** — In bioinformatics, the root mean square deviation of atomic positions, or simply root mean square deviation (RMSD), is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed molecules.
- **Protein structure prediction** — Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its secondary and tertiary structure from primary structure.
- **Critical Assessment of Prediction of Interactions** — Critical Assessment of PRediction of Interactions (CAPRI) is a community-wide experiment in modelling the molecular structure of protein complexes, otherwise known as protein–protein docking.

E.9.2 Background: amino acids, energetics and enzymes

- **Amino acid** — Amino acids are organic compounds that contain both amino and carboxylic acid functional groups. Although over 500 amino acids exist in nature, by far the most important are the 22 α-amino acids incorporated into proteins.
- **Proteinogenic amino acid** — Proteinogenic amino acids are amino acids that are incorporated biosynthetically into proteins during translation from RNA. The word "proteinogenic" means "protein creating".
- **Peptide bond** — In organic chemistry, a peptide bond is an amide type of covalent chemical bond linking two consecutive alpha-amino acids from C1 of one alpha-amino acid and N2 of another, along a peptide or protein chain.
- **Gibbs free energy** — In thermodynamics, the Gibbs free energy is a thermodynamic potential that can be used to calculate the maximum amount of work, other than pressure–volume work, that may be performed by a thermodynamically closed system at constant temperature and pressure.

- **Dissociation constant** — In chemistry, biochemistry, and pharmacology, a dissociation constant (KD) is a specific type of equilibrium constant that measures the propensity of a larger object to separate (dissociate) reversibly into smaller components, as when a complex falls apart into its component molecules, or when a salt splits up into its component ions.
- **Enzyme** — An enzyme is a biological macromolecule, usually a protein, that acts as a biological catalyst, accelerating chemical reactions without being consumed in the process. The molecules on which enzymes act are called substrates, which are converted into products.
- **Enzyme catalysis** — Enzyme catalysis is the increase in the rate of a process by an "enzyme", a biological molecule. Most enzymes are proteins, and most such processes are chemical reactions.
- **Michaelis–Menten kinetics** — In biochemistry, Michaelis–Menten kinetics, named after Leonor Michaelis and Maud Menten, is the simplest case of enzyme kinetics, applied to enzyme-catalysed reactions involving the transformation of one substrate into one product.

E.9.3 Protein structure and how it is determined

- **Protein structure** — Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. Proteins are polymers – specifically polypeptides – formed from sequences of amino acids, which are the monomers of the polymer.
- **Protein secondary structure** — Protein secondary structure is the local spatial conformation of the polypeptide backbone excluding the side chains. The two most common secondary structural elements are alpha helices and beta sheets, though beta turns and omega loops occur as well.
- **Ramachandran plot** — In biochemistry, a Ramachandran plot, originally developed in 1963 by G. N. Ramachandran, C. Ramakrishnan, and V.
- **X-ray crystallography** — X-ray crystallography is the experimental science of determining the atomic and molecular structure of a crystal, in which the structure causes a beam of incident X-rays to diffract in specific directions.
- **R-factor (crystallography)** — In crystallography, the R-factor is a measure of the disagreement between the crystallographic model and the experimental X-ray diffraction data - lower the R value lower is the disagreement or better is the agreement.
- **Protein nuclear magnetic resonance spectroscopy** — Nuclear magnetic resonance spectroscopy of proteins is a field of structural biology in which NMR spectroscopy is used to obtain information about the structure and dynamics of proteins, and also nucleic acids, and their complexes.
- **Cryogenic electron microscopy** — Cryogenic electron microscopy (cryo-EM) is a transmission electron microscopy technique applied to samples cooled to cryogenic temperatures.
- **Single particle analysis** — Single particle analysis is a group of related computerized image processing techniques used to analyze images from transmission electron microscopy (TEM).

E.9.4 Protein folding

- **Protein folding** — Protein folding is the physical process by which a protein, after synthesis by a ribosome as a linear chain of amino acids, changes from an unstable random coil into a more ordered three-dimensional structure.
- **Anfinsen's dogma** — Anfinsen's dogma, also known as the thermodynamic hypothesis, is a postulate in molecular biology.
- **Levinthal's paradox** — Levinthal's paradox is a thought experiment in the field of computational protein structure prediction that concerns the mechanism through which peptides reach a stable native configuration, given the large amount of folding conformations.

- **Folding funnel** — The folding funnel hypothesis is a specific version of the energy landscape theory of protein folding, which assumes that a protein's native state corresponds to its free energy minimum under the solution conditions usually encountered in cells.
- **Hydrophobic effect** — The hydrophobic effect is the observed tendency of nonpolar substances to aggregate in an aqueous solution and to be excluded by water.
- **Molten globule** — In molecular biology, the term molten globule (MG) refers to protein states that are more or less compact, but are lacking the specific tight packing of amino acid residues which creates the solid state-like tertiary structure of completely folded proteins.
- **Chaperone (protein)** — In molecular biology, molecular chaperones are proteins that assist the conformational folding or unfolding of proteins or macromolecular protein complexes.

E.9.5 Structure comparison and evolution

- **Structural alignment** — Structural alignment attempts to establish homology between two or more polymer structures based on their shape and three-dimensional conformation. This process is usually applied to protein tertiary structures but can also be used for large RNA molecules.
- **Structural alignment software** — This list of structural comparison and alignment software is a compilation of software tools and web portals used in pairwise or multiple structural comparison and structural alignment.
- **Root-mean-square deviation of atomic positions** — In bioinformatics, the root mean square deviation of atomic positions, or simply root mean square deviation (RMSD), is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed molecules.
- **Template modeling score** — In bioinformatics, the template modeling score or TM-score is a measure of similarity between two protein structures.
- **Protein fold** — A protein superfamily is the largest grouping (clade) of proteins for which common ancestry can be inferred. Usually this common ancestry is inferred from structural alignment and mechanistic similarity, even if no sequence similarity is evident.
- **Protein domain** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure.
- **Protein superfamily** — A protein superfamily is the largest grouping (clade) of proteins for which common ancestry can be inferred. Usually this common ancestry is inferred from structural alignment and mechanistic similarity, even if no sequence similarity is evident.

E.9.6 Protein structure databases

- **Protein Data Bank** — The Protein Data Bank (PDB) is a database for the three-dimensional structural data of large biological molecules such as proteins and nucleic acids, which is overseen by the Worldwide Protein Data Bank (wwPDB).
- **Structural Classification of Proteins database** — The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences.
- **CATH database** — The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains.
- **Evolutionary Classification of Protein Domains** — The Evolutionary Classification of Protein Domains (ECOD) is a biological database that classifies protein domains available from the Protein Data Bank. The ECOD tries to determine the evolutionary relationships between proteins.
- **AlphaFold Protein Structure Database** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.

- **Protein domain** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure.

E.9.7 Protein structure prediction

- **Protein structure prediction** — Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its secondary and tertiary structure from primary structure.
- **Homology modeling** — Homology modeling, also known as comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein.
- **Threading (protein sequence)** — In molecular biology, protein threading, also known as fold recognition, is a method of protein modeling which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure.
- **De novo protein structure prediction** — In computational biology, de novo protein structure prediction refers to an algorithmic process by which protein tertiary structure is predicted from its amino acid primary sequence.
- **Direct coupling analysis** — Direct coupling analysis or DCA is an umbrella term comprising several methods for analyzing sequence data in computational biology.
- **Protein contact map** — A protein contact map represents the distance between all possible amino acid residue pairs of a three-dimensional protein structure using a binary two-dimensional matrix.
- **AlphaFold** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.
- **Intrinsically disordered proteins** — In molecular biology, an intrinsically disordered protein (IDP) is a protein that lacks a fixed or ordered three-dimensional structure, typically in the absence of its macromolecular interaction partners, such as other proteins or RNA.

E.9.8 Protein language models and structure

- **Large language model** — A large language model (LLM) is an AI model trained on a vast amount of text for natural language processing tasks, especially language generation. LLMs can typically generate, summarize, translate, and analyze text in many contexts.
- **Transformer (deep learning architecture)** — In deep learning, the transformer is a family of artificial neural network architectures based on the multi-head attention mechanism, in which input data such as text, images, or audio is converted to a sequence of numerical representations called tokens, and each token is converted into a vector through an embedding layer.
- **BERT (language model)** — Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google. It learns to represent text as a sequence of vectors using self-supervised learning.
- **Word embedding** — In natural language processing, a word embedding is a representation of a word. The embedding is used in text analysis.
- **Multiple sequence alignment** — Multiple sequence alignment (MSA) is the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.
- **Metagenomics** — Metagenomics is the study of all genetic material from all organisms in a particular environment, providing insights into their composition, diversity, and functional potential.

- **AlphaFold** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.

E.9.9 Membrane proteins

- **Membrane protein** — Membrane proteins are common proteins that are part of, or interact with, biological membranes. Membrane proteins fall into several broad categories depending on their location.
- **Transmembrane protein** — A transmembrane protein is a type of integral membrane protein that spans the entirety of the cell membrane. Many transmembrane proteins function as gateways to permit the transport of specific substances across the membrane.
- **Transmembrane domain** — A transmembrane domain is a membrane-spanning protein domain. TMDs may consist of one or several alpha-helices or a transmembrane beta barrel.
- **Membrane topology** — Topology of a transmembrane protein refers to locations of N- and C-termini of membrane-spanning polypeptide chain with respect to the inner or outer sides of the biological membrane occupied by the protein.
- **Hydrophobicity scales** — Hydrophobicity scales are values that define the relative hydrophobicity or hydrophilicity of amino acid residues. The more positive the value, the more hydrophobic are the amino acids located in that region of the protein.
- **Translocon** — The translocon is a general term for a protein channel in biological membranes that functions to move polypeptides across the membrane or insert them into the lipid bilayer.
- **Beta barrel** — In protein structures, a beta barrel (β barrel) is a beta sheet composed of tandem repeats that twists and coils to form a closed toroidal structure in which the first strand is bonded to the last strand.
- **Bacterial outer membrane** — The bacterial outer membrane is found in gram-negative bacteria. Gram-negative bacteria form two lipid bilayers in their cell envelopes - an inner membrane (IM) that encapsulates the cytoplasm, and an outer membrane (OM) that encapsulates the periplasm.

E.9.10 Dynamics and docking

- **Molecular dynamics** — Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system.
- **Force field (chemistry)** — In the context of chemistry, molecular physics, physical chemistry, and molecular modelling, a force field is a computational model that is used to describe the forces between atoms within molecules or between molecules as well as in crystals.
- **Protein dynamics** — In molecular biology, proteins are generally thought to adopt unique structures determined by their amino acid sequences. However, proteins are not strictly static objects, but rather populate ensembles of conformations.
- **Normal mode** — A normal mode of a dynamical system is a pattern of motion in which all parts of the system move sinusoidally with the same frequency and with a fixed phase relation. The free motion described by the normal modes takes place at fixed frequencies.
- **Anisotropic Network Model** — The Anisotropic Network Model (ANM) is a simple yet powerful tool made for normal mode analysis of proteins, which has been successfully applied for exploring the relation between function and dynamics for many proteins.
- **Gaussian network model** — The Gaussian network model (GNM) is a representation of a biological macromolecule as an elastic mass-and-spring network to study, understand, and characterize the mechanical aspects of its long-time large-scale dynamics.

- **Docking (molecular)** — In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when a ligand and a target are bound to each other to form a stable complex.
- **Scoring functions for docking** — In the fields of computational chemistry and molecular modelling, scoring functions are mathematical functions used to approximately predict the binding affinity between two molecules after they have been docked.

E.9.11 Protein design

- **Protein design** — Protein design is the rational design of new protein molecules to design novel activity, behavior, or purpose, and to advance basic understanding of protein function.
- **Protein engineering** — Protein engineering is the process of developing useful or valuable proteins through the design and production of unnatural polypeptides, often by altering amino acid sequences found in nature.
- **Diffusion model** — In machine learning, diffusion models, also known as diffusion-based generative models or score-based generative models, are a class of latent variable generative models.
- **Rosetta@home** — Rosetta@home is a volunteer computing project researching protein structure prediction on the Berkeley Open Infrastructure for Network Computing (BOINC) platform, run by the Baker lab.
- **Protein folding** — Protein folding is the physical process by which a protein, after synthesis by a ribosome as a linear chain of amino acids, changes from an unstable random coil into a more ordered three-dimensional structure.

E.10 Protein function

Wikipedia articles for Chapter 10.

E.10.1 Definitions of function

- **Gene Ontology** — The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
- **Ontology (information science)** — In information science, an ontology encompasses a representation, formal naming, and definitions of the categories, properties, and relations between the concepts, data, or entities that pertain to one, many, or all domains of discourse.
- **Enzyme Commission number** — The Enzyme Commission number is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze.
- **Enzyme** — An enzyme is a biological macromolecule, usually a protein, that acts as a biological catalyst, accelerating chemical reactions without being consumed in the process. The molecules on which enzymes act are called substrates, which are converted into products.
- **Moonlighting protein** — Protein moonlighting is a phenomenon by which a protein can perform more than one function. It is an example of gene sharing.
- **Protein function prediction** — Protein function prediction methods are techniques that bioinformatics researchers use to assign biological or biochemical roles to proteins. These proteins are usually ones that are poorly studied or predicted based on genomic sequence data.

E.10.2 Function prediction

- **Protein function prediction** — Protein function prediction methods are techniques that bioinformatics researchers use to assign biological or biochemical roles to proteins. These proteins are usually ones that are poorly studied or predicted based on genomic sequence data.
- **Sequence homology** — Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life.
- **Homology (biology)** — In biology, homology is similarity in anatomical structures or genes between organisms of different taxa due to shared ancestry, regardless of current functional differences.
- **Protein domain** — In molecular biology, a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and folds independently from the rest. Each domain forms a compact, folded three-dimensional structure.
- **InterPro** — InterPro is a database of protein families, protein domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences in order to functionally characterise them.
- **Phylogenetic profiling** — Phylogenetic profiling is a bioinformatics technique in which the joint presence or joint absence of two traits across large numbers of species is used to infer a meaningful biological connection, such as involvement of two different proteins in the same biological pathway.
- **Operon** — In genetics, an operon is a functioning unit of DNA containing a cluster of genes under the control of a single promoter.
- **Short linear motif** — In molecular biology short linear motifs (SLiMs), linear motifs or minimotifs are short stretches of protein sequence that mediate protein–protein interaction.

E.10.3 Subcellular localisation

- **Protein targeting** — Protein targeting or protein sorting is the biological mechanism by which proteins are transported to their appropriate destinations within or outside the cell.
- **Signal peptide** — A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined toward the secretory pathway.
- **Signal recognition particle** — The signal recognition particle (SRP) is an abundant, cytosolic, universally conserved ribonucleoprotein that recognizes and targets specific proteins to the endoplasmic reticulum in eukaryotes and the plasma membrane in prokaryotes.
- **Transit peptide** — A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined toward the secretory pathway.
- **Mitochondrial targeting sequence** — A target peptide is a short peptide consensus sequence that directs the transport of a protein to a specific region in the cell, including the nucleus, mitochondria, endoplasmic reticulum (ER), chloroplast, apoplast, peroxisome and plasma membrane.
- **Nuclear localization sequence** — A nuclear localization sequence, or signal (NLS) is an amino acid sequence motif that 'tags' a protein for import into the cell nucleus by nuclear transport.
- **Protein subcellular localization prediction** — Protein subcellular localization prediction involves the prediction of where a protein resides in a cell, its subcellular localization.

E.10.4 Predicting the effects of mutations

- **Missense mutation** — In genetics, a missense mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. It is a type of nonsynonymous substitution.

- **Multiplexed Assays of Variant Effect** — Multiplexed Assays of Variant Effect (MAVE) is a method used to determine sequence-function relationships, identify variants that alter protein structure, stability, or activity, and characterize regulatory elements.
- **ClinVar** — ClinVar is a public archive with free access to reports on the relationships between human genetic variations and phenotypes, with supporting evidence. The database includes germline and somatic variants of any size, type or genomic location.
- **Variant of uncertain significance** — A variant of uncertain significance (VUS) is a genetic variant that has been identified through genetic testing but whose significance to the function or health of an organism is not known.

E.10.5 Benchmarking function prediction

- **Critical Assessment of Function Annotation** — The Critical Assessment of Function Annotation (CAFA) is an ongoing community-driven experiment designed to evaluate computational methods for protein function prediction.
- **Protein function prediction** — Protein function prediction methods are techniques that bioinformatics researchers use to assign biological or biochemical roles to proteins. These proteins are usually ones that are poorly studied or predicted based on genomic sequence data.
- **Precision and recall** — In classification, precision and recall are performance metrics that apply to data retrieved from a collection, corpus or sample space.
- **Open-world assumption** — The closed-world assumption (CWA), in a formal system of logic used for knowledge representation, is the presumption that a statement that is true is also known to be true. Therefore, by the contrapositive, what is not currently known to be true must be false.

E.10.6 Experimental determination of function

- **Forward genetics** — Forward genetics is a molecular genetics approach of determining the genetic basis responsible for a phenotype.
- **Reverse genetics** — Reverse genetics is a method in molecular genetics that is used to help understand the function(s) of a gene by analysing the phenotypic effects caused by genetically engineering specific nucleic acid sequences within the gene.
- **CRISPR gene editing** — CRISPR gene editing is a genetic engineering technique in molecular biology by which the genomes of living organisms may be modified. It is based on a simplified version of the bacterial CRISPR-Cas9 antiviral defense system.
- **Genetic screen** — A genetic screen or mutagenesis screen is an experimental technique used to identify and select individuals who possess a phenotype of interest in a mutagenized population. Hence a genetic screen is a type of phenotypic screen.
- **Gene knockout** — Gene knockouts are a widely used genetic engineering technique that involves the targeted removal or inactivation of a specific gene within an organism's genome.
- **Essential gene** — Essential genes are indispensable genes for organisms to grow and reproduce offspring under certain environment. However, being essential is highly dependent on the circumstances in which an organism lives.

E.11 Protein–protein interactions

Wikipedia articles for Chapter 11.

E.11.1 Experimental methods

- **Protein–protein interaction** — Protein–protein interactions (PPIs) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect.
- **Two-hybrid screening** — Two-hybrid screening is a molecular biology technique used to discover protein–protein interactions (PPIs) and protein–DNA interactions by testing for physical interactions between two proteins or a single protein and a DNA molecule, respectively.
- **Tandem affinity purification** — Tandem affinity purification (TAP) is an immunoprecipitation-based purification technique for studying protein–protein interactions. The goal is to extract from a cell only the protein of interest, in complex with any other proteins it interacted with.
- **Co-immunoprecipitation** — Immunoprecipitation (IP) is the technique of precipitating a protein antigen out of solution using an antibody that specifically binds to that particular protein.
- **Protein-fragment complementation assay** — Within the field of molecular biology, a protein-fragment complementation assay, or PCA, is a method for the identification and quantification of protein–protein interactions.
- **Synthetic lethality** — Synthetic lethality is defined as a type of genetic interaction where the combination of two genetic events results in cell death or death of an organism.
- **Epistasis** — Epistasis is a phenomenon in genetics in which the effect of a mutation on a trait is dependent on the presence or absence of mutations at different loci, that may be on the same gene (intragenic) or on different genes, called modifier genes.
- **Gene regulatory network** — A gene regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins which, in turn, determine the function of the cell.

E.11.2 Networks

- **Interactome** — In molecular biology, an interactome is the whole set of molecular interactions in a particular cell. The term specifically refers to physical interactions among molecules but can also describe sets of indirect interactions among genes.
- **Biological network** — A biological network is a method of representing systems as complex sets of binary interactions or relations between various biological entities. In general, networks or graphs are used to capture relationships between entities or objects.
- **Scale-free network** — A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction P(k) of nodes in the network having k connections to other nodes goes for large values of k as
- **Small-world network** — A small-world network is a graph characterized by a high clustering coefficient and low distances. In an example of a social network, high clustering implies the high probability that two friends of one person are friends themselves.
- **Clustering coefficient** — In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.
- **Network motif** — Network motifs are recurrent and statistically significant subgraphs or patterns of a larger graph.
- **STRING** — In molecular biology, STRING is a biological database and web resource of known and predicted protein–protein interactions.
- **Network medicine** — Network medicine is the application of network science towards identifying, preventing, and treating diseases. This field focuses on using network topology and network dynamics towards identifying diseases and developing medical drugs.
- **Systems biology** — Systems biology is the computational and mathematical analysis and modeling of complex biological systems.

- **Flux balance analysis** — In biochemistry and systems biology, flux balance analysis (FBA) is a mathematical method for simulating the metabolism of cells or entire unicellular organisms, such as E. coli or yeast, using genome-scale reconstructions of metabolic networks.

E.11.3 Prediction methods

- **Protein–protein interaction prediction** — Protein–protein interaction prediction is a field combining bioinformatics and structural biology in an attempt to identify and catalog physical interactions between pairs or groups of proteins.
- **Phylogenetic profiling** — Phylogenetic profiling is a bioinformatics technique in which the joint presence or joint absence of two traits across large numbers of species is used to infer a meaningful biological connection, such as involvement of two different proteins in the same biological pathway.
- **Direct coupling analysis** — Direct coupling analysis or DCA is an umbrella term comprising several methods for analyzing sequence data in computational biology.
- **Docking (molecular)** — In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when a ligand and a target are bound to each other to form a stable complex.
- **Protein–protein docking** — Macromolecular docking is the computational modelling of the quaternary structure of complexes formed by two or more interacting biological macromolecules.
- **Critical Assessment of Prediction of Interactions** — Critical Assessment of PRediction of Interactions (CAPRI) is a community-wide experiment in modelling the molecular structure of protein complexes, otherwise known as protein–protein docking.
- **AlphaFold** — AlphaFold is an artificial intelligence (AI) program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure. It is designed using deep learning techniques.
- **Biomolecular condensate** — In biochemistry, biomolecular condensates are a class of membraneless organelles and organelle subdomains, which carry out specialized functions within the cell.

E.12 Machine learning

Wikipedia articles for Chapter 12.

E.12.1 Basics

- **Machine learning** — Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without being explicitly programmed.
- **Overfitting** — In mathematical modeling, overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit to additional data or predict future observations reliably.
- **Bias–variance tradeoff** — In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.
- **Cross-validation (statistics)** — Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
- **Leakage (machine learning)** — In statistics and machine learning, leakage refers to the use of information during model training that would not be available at prediction time.

- **Precision and recall** — In classification, precision and recall are performance metrics that apply to data retrieved from a collection, corpus or sample space.
- **Receiver operating characteristic** — A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier model at varying threshold values.
- **Phi coefficient** — In statistics, the phi coefficient, also known as the mean square contingency coefficient or Yule coefficient of correlation and commonly denoted by φ or $r\varphi$, is a measure of association between two binary variables.

E.12.2 A short history of machine learning in bioinformatics

- **Perceptron** — In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function that can decide whether or not an input, represented by a vector of numbers, belongs to some specific class.
- **Artificial neural network** — In machine learning, a neural network (NN) or artificial neural network (ANN) is a computational model inspired by the structure and functions of biological neural networks.
- **Backpropagation** — In machine learning, backpropagation is a gradient computation method commonly used for training a neural network in computing parameter updates.
- **Hidden Markov model** — In probability theory, a hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process. An HMM requires that there be an observable process whose outcomes depend on the outcomes of in a known way.
- **Support vector machine** — In machine learning, a support vector machine (SVM) or support vector network is a supervised max-margin model with associated learning algorithms that analyze data for classification and regression analysis.
- **Kernel method** — In machine learning, kernel machines are a class of algorithms for pattern analysis, whose best known member is the support-vector machine (SVM). These methods involve using linear classifiers to solve nonlinear problems.
- **Random forest** — Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training.
- **Gradient boosting** — Gradient boosting is a machine learning technique based on boosting in a functional space, where the target is pseudo-residuals instead of residuals as in traditional boosting.

E.12.3 Deep learning

- **Deep learning** — In machine learning, deep learning (DL) focuses on utilizing multilayered neural networks to perform tasks such as classification, regression, and representation learning.
- **Stochastic gradient descent** — Stochastic gradient descent is an iterative method for optimizing an objective function with suitable smoothness properties.
- **Convolutional neural network** — A convolutional neural network (CNN) is a type of feedforward neural network that learns features via filter optimization.
- **Recurrent neural network** — In artificial neural networks, recurrent neural networks (RNNs) are designed for processing sequential data, such as text, speech, and time series, where the order of elements is important.
- **Long short-term memory** — Long short-term memory (LSTM) is a type of recurrent neural network (RNN) aimed at mitigating the vanishing gradient problem commonly encountered by traditional RNNs.

- **Transformer (deep learning architecture)** — In deep learning, the transformer is a family of artificial neural network architectures based on the multi-head attention mechanism, in which input data such as text, images, or audio is converted to a sequence of numerical representations called tokens, and each token is converted into a vector through an embedding layer.
- **Graph neural network** — Graph neural networks (GNNs) are artificial neural networks designed for tasks whose inputs are graphs.
- **Explainable artificial intelligence** — Within artificial intelligence (AI), explainable AI (XAI), generally overlapping with interpretable AI, interpretable machine learning and explainable machine learning (XML), is a field of research that explores methods that provide humans with the ability of intellectual oversight over AI algorithms.

E.12.4 Generative models

- **Generative model** — Generative models are a class of computational models frequently used for classification.
- **Autoregressive model** — In statistics, an autoregressive (AR) model is a modelled representation of a type of random process. It can be used to describe time-varying processes from many natural and artificial sources.
- **Autoencoder** — An autoencoder is a type of artificial neural network used to learn efficient codings of unlabeled data.
- **Variational autoencoder** — In machine learning, a variational autoencoder (VAE) is an artificial neural network architecture introduced by Diederik P. Kingma and Max Welling in 2013. It is part of the families of probabilistic graphical models and variational Bayesian methods.
- **Generative adversarial network** — A generative adversarial network (GAN) is a class of machine learning frameworks and a prominent framework for approaching generative AI. The concept was initially developed by Ian Goodfellow and his colleagues at Université de Montréal in June 2014.
- **Diffusion model** — In machine learning, diffusion models, also known as diffusion-based generative models or score-based generative models, are a class of latent variable generative models.

E.12.5 Foundation models

- **Foundation model** — In artificial intelligence, a foundation model (FM), also known as large x model, is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases.
- **Self-supervised learning** — Self-supervised learning (SSL) is a paradigm in machine learning where a model is trained on a task using the data itself to generate supervisory signals, rather than relying on externally-provided labels.
- **Transfer learning** — Transfer learning (TL) is a technique in machine learning (ML) in which knowledge learned from a task is re-used in order to boost performance on a related task.
- **Fine-tuning (deep learning)** — In deep learning, fine-tuning is the process of adapting a computational model trained for one task to perform a different, usually more specific, task.
- **Neural scaling law** — In machine learning, a neural scaling law is an empirical scaling law that describes how neural network performance changes as key factors are scaled up or down.
- **Retrieval-augmented generation** — Retrieval-augmented generation (RAG) is a technique that enables large language models (LLMs) to retrieve and incorporate new information from external data sources. With RAG, LLMs first refer to a specified set of documents, then respond to user queries.

E.12.6 AI in research practice

- **Artificial intelligence** — Artificial intelligence (AI) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making.
- **Large language model** — A large language model (LLM) is an AI model trained on a vast amount of text for natural language processing tasks, especially language generation. LLMs can typically generate, summarize, translate, and analyze text in many contexts.
- **Generative artificial intelligence** — Generative artificial intelligence (GenAI) is a subfield of artificial intelligence (AI) that uses generative models to generate text, images, videos, audio, computer code or other forms of digital data.
- **Reinforcement learning from human feedback** — In machine learning, reinforcement learning from human feedback (RLHF) is a technique to align an intelligent agent with human preferences.
- **Hallucination (artificial intelligence)** — In the field of artificial intelligence (AI), a hallucination or artificial hallucination is a response generated by AI that contains false or misleading information presented as fact.
- **Ethics of artificial intelligence** — The ethics of artificial intelligence covers a broad range of topics within AI that are considered to have particular ethical stakes.

E.12.7 Integrating data

- **Multiomics** — Multiomics, multi-omics, integrative omics, "panomics" or "pan-omics" is a biological analysis approach in which the data consists of multiple "omes", such as the genome, epigenome, transcriptome, proteome, metabolome, exposome, and microbiome ; in other words, the use of multiple omics technologies to study life in a concerted way.
- **Data integration** — Data integration is the process of combining, sharing, or synchronizing data from multiple sources to provide users with a unified view. There are a wide range of possible applications for data integration, from commercial to scientific.
- **Canonical correlation** — In statistics, canonical-correlation analysis (CCA), also called canonical variates analysis, is a way of inferring information from cross-covariance matrices.
- **Mendelian randomization** — In epidemiology, Mendelian randomization is a method using measured variation in genes to examine the causal effect of an exposure on an outcome.
- **Causal inference** — Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system.

E.13 Image analysis and models of cells

Wikipedia articles for Chapter 13.

E.13.1 Images as data

- **Microscopy** — Microscopy is the technical field of using microscopes to view subjects too small to be seen with the naked eye.
- **Fluorescence microscope** — A fluorescence microscope is an optical microscope that uses fluorescence instead of, or in addition to scattering, reflection, and attenuation or absorption, to study the properties of organic or inorganic substances.
- **Confocal microscopy** — Confocal microscopy is an optical imaging technique for increasing optical resolution and contrast of a micrograph by means of using a spatial pinhole to block out-of-focus light in image formation.

- **Image segmentation** — In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple image segments, also known as image regions or image objects.
- **Watershed (image processing)** — In the study of image processing, a watershed is a transformation defined on a grayscale image. The name refers metaphorically to a geological watershed, or drainage divide, which separates adjacent drainage basins.
- **Fiji (software)** — Fiji is an open source image processing package based on ImageJ2.
- **CellProfiler** — CellProfiler is free, open-source software designed to enable biologists without training in computer vision or programming to quantitatively measure phenotypes from thousands of images automatically.

E.13.2 Deep learning for images

- **Convolutional neural network** — A convolutional neural network (CNN) is a type of feedforward neural network that learns features via filter optimization.
- **U-Net** — U-Net is a convolutional neural network that was developed for image segmentation.
- **Vision transformer** — A vision transformer (ViT) is a transformer designed for computer vision. A ViT decomposes an input image into a series of patches, serializes each patch into a vector, and maps it to a smaller dimension with a single matrix multiplication.
- **Self-supervised learning** — Self-supervised learning (SSL) is a paradigm in machine learning where a model is trained on a task using the data itself to generate supervisory signals, rather than relying on externally-provided labels.
- **Segment Anything** — Segment Anything is a series of open-source computer vision foundation models developed by Meta AI for image and video segmentation.
- **ImageNet** — The ImageNet project is a large visual database designed for use in visual object recognition software research.

E.13.3 Image-based profiling

- **High-content screening** — High-content screening (HCS), also known as high-content analysis (HCA) or cellomics, is a method that is used in biological research and drug discovery to identify substances such as small molecules, peptides, or RNAi that alter the phenotype of a cell in a desired manner.
- **Cell Painting** — The cell painting assay is a high-content, high-throughput imaging technique used to capture a wide array of cellular phenotypes in response to diverse perturbations.
- **Phenotypic screening** — Phenotypic screening is a type of screening used in biological research and drug discovery to identify substances such as small molecules, peptides, or RNAi that alter the phenotype of a cell or an organism in a desired manner.
- **Mechanism of action** — In pharmacology, the term mechanism of action (MOA) refers to the specific biochemical interaction through which a drug substance produces its pharmacological effect.

E.13.4 Protein localisation from images

- **Human Protein Atlas** — The Human Protein Atlas (HPA) is a Swedish-based program started in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

- **Immunofluorescence** — Immunofluorescence (IF) is a light microscopy-based technique that allows detection and localization of a wide variety of target biomolecules within a cell or tissue at a quantitative level.
- **Subcellular localization** — The cells of eukaryotic organisms are elaborately subdivided into functionally-distinct membrane-bound compartments.
- **Green fluorescent protein** — The green fluorescent protein (GFP) is a protein that exhibits green fluorescence when exposed to light in the blue to ultraviolet range.
- **Protein tag** — Protein tags are peptide sequences genetically grafted onto a recombinant protein. Tags are attached to proteins for various purposes.

E.13.5 Tissues: multiplexed imaging and computational pathology

- **Digital pathology** — Digital pathology is a sub-field of pathology that focuses on managing and analyzing information generated from digitized specimen slides.
- **Spatial transcriptomics** — Spatial transcriptomics, or spatially resolved transcriptomics, is a method that captures positional context of transcriptional activity within intact tissue.
- **Mass cytometry** — Mass cytometry is a high-dimensional single-cell analysis technique that integrates flow cytometry with time of flight mass spectrometry. It is used for the determination of the properties of cells (cytometry).
- **Multiple instance learning** — In machine learning, multiple-instance learning (MIL) is a type of supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances.
- **Histopathology** — Histopathology is the microscopic examination of tissue in order to study the manifestations of disease.
- **H&E stain** — Hematoxylin and eosin stain is one of the principal tissue stains used in histology. It is the most widely used stain in medical diagnosis and is often the gold standard.

E.13.6 Cell atlases and virtual cells

- **Human Cell Atlas** — The Human Cell Atlas is a global project to describe all cell types in the human body, to understand human health and for diagnosing, monitoring, and treating disease.
- **Virtual cell** — A cellular model or virtual cell is a computational model of aspects of a biological cell, for the purposes of in silico research.
- **Foundation model** — In artificial intelligence, a foundation model (FM), also known as large x model, is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases.
- **Perturb-seq** — Perturb-seq refers to a high-throughput method of performing single cell RNA sequencing (scRNA-seq) on pooled genetic perturbation screens.
- **Chan Zuckerberg Initiative** — The Chan Zuckerberg Initiative (CZI) is an organization established and owned by Facebook founder Mark Zuckerberg and his wife Priscilla Chan with an investment of 99 percent of the couple's wealth from their Facebook shares over their lifetime.

E.13.7 Whole-cell models and cryo-electron tomography

- **Cryogenic electron tomography** — Cryogenic electron tomography (cryoET) is an imaging technique used to reconstruct high-resolution (~1–4 nm) three-dimensional volumes of samples, often biological macromolecules and cells. cryoET is a specialized application of transmission electron cryomicroscopy (CryoTEM) in which samples are imaged as they are tilted, resulting in a series of 2D images that can be combined to produce a 3D reconstruction, similar to a CT scan of the human body.

- **Mycoplasma genitalium** — Mycoplasma genitalium is a sexually transmitted, small pathogenic bacterium that lives on the mucous epithelial cells of the urinary and genital tracts in humans. Medical journals in 2007 and 2015 reported that Mgen had become more common.
- **Minimal genome** — Minimal genome is the theoretical smallest set of genes sufficient for life to exist and propagate under nutrient-rich and stress-free conditions.
- **Molecular dynamics** — Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system.
- **Systems biology** — Systems biology is the computational and mathematical analysis and modeling of complex biological systems.

F Cross-topic problems and case studies

> **Credit**
>
> The problems and case studies in this appendix are adapted from Matteo Tadiello, *Modern Bioinformatics: From Foundations to Frontiers* (CC BY 4.0), used with permission.

The review problems at the end of each chapter test one topic at a time. In the oral exam, questions often move from one topic to another: from a sequence search to evolution, from structure to function, from an experiment to its statistics. The problems in this appendix combine several chapters and have no single correct answer. Use them to practise explaining a complete line of reasoning: what the biological question is, which data and methods would address it, what each result can and cannot show, and how the conclusion would be validated.

F.1 Cross-topic problems

These problems integrate concepts from at least three chapters. For each problem, define the biological claim, identify the data-generating process, justify the computational formulation, state assumptions, compare alternatives, and propose validation.

1. You discover a previously unannotated bacterial gene associated with antibiotic resistance. Design the sequence, structure, evolutionary, expression, and experimental workflow needed before making a functional claim.
2. A profile HMM and a structure search disagree about a remote homolog. Explain possible causes and how you would adjudicate them.
3. A gene is differentially expressed but its protein abundance is unchanged. Give technical and biological explanations and design follow-up tests.
4. A pangenome graph improves variant recall. Explain how you would demonstrate that the gain is real rather than benchmark leakage or truth-set bias.
5. A high-confidence predicted protein complex has no AP-MS support. Reconcile the evidence and propose orthogonal validation.
6. A gene family tree conflicts with the accepted species tree. Compare duplication/loss, incomplete lineage sorting, horizontal transfer, recombination, and artefact explanations.
7. A classifier predicts disease from transcriptomes with AUROC 0.97 but fails prospectively. Diagnose data leakage, batch structure, spectrum effects, calibration, and target shift.
8. An RNA-seq study has three treatment samples processed on one day and three controls on another. Explain what can and cannot be recovered computationally.
9. A protein has 35% sequence identity to a characterised enzyme. Build an evidence hierarchy for transferring catalytic function and substrate specificity.
10. Compare the meaning of an E-value, adjusted p-value, posterior probability, bootstrap proportion, predicted confidence score, and calibrated class probability.
11. Design a database record that distinguishes raw evidence, inferred annotation, version, uncertainty, and provenance.

12. Explain how repeats affect read mapping, genome assembly, gene annotation, expression quantification, and evolutionary inference.
13. A rare cell population appears only after single-cell integration. Explain how to decide whether it is biology or an integration artefact.
14. You need to identify direct interaction partners of a membrane protein in living cells. Compare experimental and computational options.
15. A foundation model produces accurate embeddings. Explain what benchmarks would establish biological utility without circularity.
16. Design a reproducible multi-omics study from sampling through public deposition, including negative controls and validation.

F.2 Case study A: from sequence to mechanism

This case study traces a biological question from comparative genomics through molecular mechanism. The routes can be analysed sequentially or assigned as separate problem sets.

F.2.1 The problem

A conserved protein of unknown function is found in several pathogenic bacteria but is absent from closely related non-pathogens. Design a strategy to determine its origin, molecular function, structure, interaction partners, and contribution to pathogenicity.

F.2.1.1 Route 1: databases and sequence search

- Which databases do you query, and how do you preserve identifier/version provenance?
- When would you choose BLAST, a profile HMM, MMseqs2, or structure search?
- How do E-values change with search space and multiple queries?
- What would make an apparent homolog a false lead?

F.2.1.2 Route 2: evolution and phylogeny

- How would you distinguish vertical inheritance, loss, and horizontal transfer?
- What alignment and substitution-model problems could distort the tree?
- How do you reconcile a gene tree with a species tree?
- What evidence would support selection without overinterpreting dN/dS?

F.2.1.3 Route 3: structure and function

- What does a high-confidence predicted fold establish?
- How would you evaluate a putative active site or ligand-binding pocket?
- When is docking informative, and how would you validate it?
- How would disorder, oligomerisation, or membrane localisation change the plan?

F.2.1.4 Route 4: expression and interactions

- Design a condition-aware expression experiment with biological replication.
- Choose an interaction assay and explain its false-positive/negative profile.
- What would count as orthogonal evidence for a direct interaction?

F.2.1.5 Route 5: integration

- Build the strongest defensible conclusion from mixed evidence.
- Identify which claims remain predictive rather than causal.
- Define the minimum experiments needed before proposing a drug target.

F.2.2 Evaluation criteria

A rigorous analysis connects each biological claim to the evidence generated by the corresponding method, states algorithmic assumptions, distinguishes estimated confidence from correctness, and proposes independent validation. A tool list without an explicit inferential argument is insufficient.

F.3 Case study B: from cohort to biological claim

This case study develops an end-to-end analysis of a heterogeneous clinical cohort, from experimental design and data generation to prediction, causal interpretation, and prospective validation.

F.3.1 The problem

A cohort study collects genomes, bulk RNA-seq, single-cell RNA-seq, and plasma proteomics from patients and controls across three hospitals. The goal is to discover a disease mechanism and a predictive biomarker. Design the analysis and justify each inferential step.

F.3.1.1 Route 1: design and data generation

- What replication and randomisation are required?
- Which variables must be balanced or recorded?
- How do sequencing and mass-spectrometry error processes differ?
- What quality controls detect sample swaps and contamination?

F.3.1.2 Route 2: genomic analysis

- Choose mapping/reference representations and variant-calling strategies.
- Explain reference bias, population structure, phasing, and structural variants.
- How would you validate a candidate regulatory variant?

F.3.1.3 Route 3: expression and proteomics

- Specify the bulk RNA-seq design matrix and contrasts.
- Explain single-cell pseudoreplication and a defensible differential-expression unit.
- Describe peptide-spectrum matching, FDR, protein inference, and missing values.
- How would you interpret RNA–protein disagreement?

F.3.1.4 Route 4: machine learning

- Define the unit of splitting and prevent patient/site leakage.
- Choose metrics for a rare disease and assess calibration.
- Compare L1, L2, tree-based, and representation-learning approaches.
- How would you perform nested model selection and external validation?

F.3.1.5 Route 5: integration and causality

- Compare early, intermediate, and late multi-omics integration.
- Separate a predictive biomarker from a causal mechanism.
- Explain how site, ancestry, treatment, and cell composition could confound the result.
- Specify an orthogonal experiment and a prospective validation study.

F.3.2 Evaluation criteria

The answer must preserve the experimental unit, connect data-generating processes to statistical models, prevent leakage, control multiplicity, distinguish prediction from causation, and specify independent validation.